Modelling Aggregation Motivated Interactions in Descriptive Text Generation

Hua Cheng

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
Division of Informatics
University of Edinburgh
2002

Abstract

Aggregation as an indispensable part of a natural language generation (NLG) system is attracting more and more attention in the generation community. Because of the difficulty of the problem and the easiness to implement a simplified version which can greatly improve the quality of the generated text, little research goes into the problem in enough breadth and depth. As a result, although most, if not all, NLG systems have an aggregation component, aggregation as a research area is still at a rather primitive stage. There is no consensus on almost any aspect of the problem, including its definition, classification or location in a generation system, etc.

When we look closely at the small amount of previous research focusing on aggregation, we find that most of it treats aggregation as a process between text planning and sentence realisation (i.e., in the sentence planner) to remove redundancy in the text plan. This approach successfully modularises the problem, but fails to capture the complex relation between aggregation and other text generation tasks.

This thesis aims at revealing the complex interactions between aggregation and other generation tasks, which have not been explored by previous research, and discusses what these interactions imply for generation architectures. To study these problems in detail, the thesis focuses on embedding phenomena in descriptive text. It identifies regularities in the way human authors produce complex NPs using embedding and the interactions between embedding and such processes as document structuring and referring expression generation.

These findings motivate a set of preferences among coherence features, which capture the complex interactions that have been discovered. The preferences mainly include features of entity-based and relation-based coherence and embedding. They are implemented in ILEX-TS and GA-plan, which represent two dramatically different text planning architectures (a pipeline and a non-pipeline architecture), and the behaviours of the two systems are compared.

The thesis quantitatively evaluates the observed embedding rules using an annotated corpus and the output of the two generation systems using human judgement. It also makes an attempt to automatically evaluate the readability of a text.

Based on the results of evaluation, the thesis is able to make a number of assertions: firstly, the effect of embedding on the planning of entity-based and relation-based coherence demands it to be taken into account in text planning to affect the structuring of content; secondly, to produce a coherent text, it is important to capture the interactions between generation tasks and this should ideally be done in a better way than presented in current NLG systems; and finally, it is possible to capture the preferences among coherence features in a non-sequential way. These form the main contributions of the thesis.

Acknowledgements

I cannot imagine finishing this thesis without the help from many people. First of all, my special thanks are to my supervisors, Dr. Chris Mellish and Dr. Mick O'Donnell, for not only taking me into the area of Natural Language Generation but also constantly pushing me forward during my PhD. They are always quick in giving feedbacks and all kinds of support. The progress that I have made in both research and English would not have been possible without their tremendous insights and patience.

My thanks are then to Dr. Massimo Poesio and Dr. Renate Henschel, who took me into the GNOME project and initiated many interesting discussions. My work in GNOME compensates greatly to various aspects of this thesis.

I thank my examiners, Dr. Johanna Moore and Dr. Ehud Reiter, for their detailed comments, which have substantially improved the quality of this thesis.

My research benefits from the discussions with some established members in the field, in particular, Dr. Alistair Knott, who offered many valuable suggestions to the various issues discussed in this thesis, Dr. Richard Cox, Dr. Simona Teufel, Dr. Frank Keller and Mark Pearson, who gave advice on empirical analysis, evaluation and the design of the questionnaires, and Dr. Jean Carletta, from whom I got better ideas about using statistics.

I would also like to thank the folks who helped to fill in the questionnaires: Dr. Finlay Smith, Ian Miguel, Dr. Stephen Cresswell, Paul Bailey, Dr. Steve Polyak, Carol Rennie, Mary Ellen Foster, Dr. Mark Core, Dr. Bill Teahan, Chris McGiffen, Angus MacLean, Dr. Robin Boswell, Sandra Williams and many others.

I acknowledge the financial support of a University of Edinburgh Studentship and a British ORS award, which sponsored my research and various opportunities for attending conferences, from which I received invaluable feedback about my work.

Finally, I would like to dedicate this thesis to my family in China, especially mum and dad, and my husband Daqing for their moral support during my PhD.

Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research.

Hua Cheng
Edinburgh
February 7, 2002

Contents

List of Figures

Chapter 1

Introduction

This chapter gives an overview of the main concerns of the thesis – problems that need to be addressed to achieve desirable aggregation results. The overview starts with an introduction to natural language generation, in particular aggregation. It then discusses those problems in detail. To study them, we choose to focus on a specific type of aggregation called embedding. Some basic concepts such as Rhetorical Structure Theory and coherence, and the domain we work on are also introduced. The main contributions and the organisation of the thesis are presented at the end of the chapter.

1.1 Aggregation in Natural Language Generation

1.1.1 Natural Language Generation

The interest in natural language generation (NLG) has grown substantially in the last two decades because of the increasing need for expressing information stored in various databases in a form that humans can understand, for example, natural language. NLG concerns the production of understandable texts in natural language forms from some underlying representation of information using computer systems. It has been used in various applications to assist human computer interaction, for example, generating patient information for doctors and playing the role of a tutor to help students to learn some scientific subjects.

An NLG task generally consists of two decisions: a strategic decision concerning *what*

to say and a tactical decision concerning *how to say* it. To be more precise, six problems can be identified in this task, as suggested in (Reiter and Dale, 1997; Mellish and Dale, 1998) (slightly different terms are used in the two papers):

Content Determination : selecting relevant facts from a large body of information in the knowledge base of an NLG system to achieve some communication goals.

Document Structuring : organising the selected information into a coherent structure, for example a hierarchical tree where a rhetorical relation is used to connect each two adjacent text spans. Elsewhere, this process is also called Text Structuring, Content Structuring, etc. We will use these as identical terms in this thesis.

Referring Expression Generation : deciding which syntactic form should be used to realise a discourse entity and which properties of this entity should be included in the referring expression to identify it.

Aggregation : combining simple representations into sentence-sized chunks. This will be elaborated in the next section.

Lexicalization : choosing suitable lexical forms to express the specified information.

Surface Realization : determining how the structured text plans from Document Structuring can be realised as grammatical natural language sentences.

Most NLG systems use a modularised pipeline architecture as described in (Reiter, 1994), which consists of three main modules and information flows from one to another with little backtracking. The three modules and their correspondence with the six problems above are:

Reiter	**Reiter, Mellish & Dale**
Content Determination	- Content Determination and Document Structuring
Sentence Planning	- Referring Expression Generation, Aggregation and Lexicalization
Surface Generation	- Surface Realization

In cognitive science, the Content Determination and Sentence Planning processes of (Reiter, 1994) are called *macroplanning* and *microplanning* respectively (Levelt, 1989).

Macroplanning is also called *text planning* in NLG. We are mainly interested in Aggregation and its relation to two other problems, Document Structuring and Referring Expression Generation. We disregard such tasks as Content Determination, Lexicalization and Surface Realization. We assume that the information to be expressed has been chosen, so all of it should be included in the generated text.

Before we get into the main issues, we need to introduce an important concept for document structuring which we will come across frequently later. It is called Rhetorical Structure Theory (RST) (Mann and Thompson, 1987b), which is "a descriptive theory of a major aspect of the organisation of natural text. It is a linguistically useful method for describing natural texts, characterising their structure primarily in terms of relations that hold between parts of the text."

Rhetorical Structure Theory can be used to represent a text structure from both a descriptive and a constructive point of view (Mann and Thompson, 1987a; Mann and Thompson, 1987c). According to RST, a natural text can be described as a hierarchical structure with a nucleus/satellite or multi-nuclear relation between each two consecutive spans of the text, as shown in Figure 1.1.

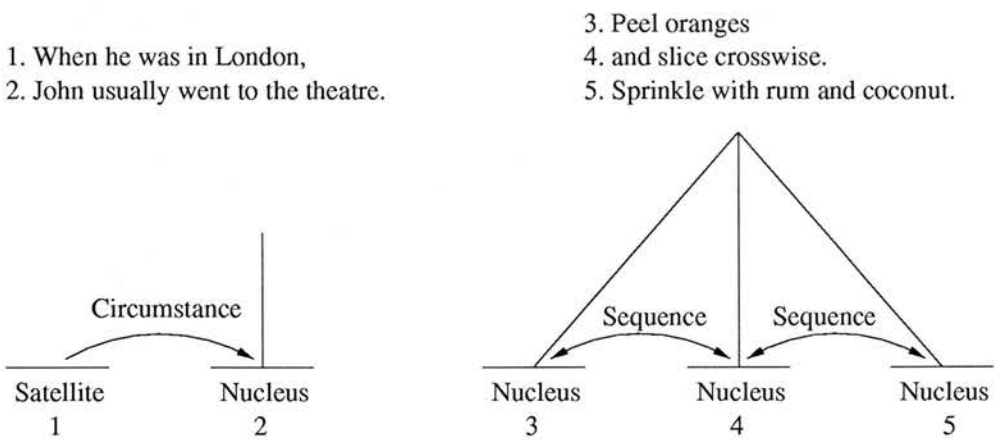


Figure 1.1: Examples of rhetorical structures

In the first example, the situation described in the satellite provides circumstances (time and location) for mentioning the situation in the nucleus; in the other example, the events mentioned by the nuclei of the SEQUENCE relation happen one after another.

Despite some problems with RST, e.g., as argued in (Moore and Pollack, 1992), it is

widely used in NLG systems. In this thesis, we take RST as a good account of at least some aspects of text coherence, and assume that a tree structure resembling an RST tree is the output of the document structuring process.

From the next section, we will focus on aggregation and introduce related issues.

1.1.2 Aggregation

Mann and Moore (1981) formulate the problem of generating text from a general purpose knowledge base as a Fragment-and-Compose paradigm, that is, first segmenting the given data structure into small manageable units, which are normally smaller than a sentence, and then combining these fragments into good sentences and paragraphs.

This paradigm is still more or less followed in the designs of current NLG systems, where simple and detailed representations of knowledge are the input to the core of the systems. Researchers often make assumptions similar to that of (Scott and de Souza, 1990), i.e., a text is to be generated, at some stage, from "verb-based, clause-sized propositions, each of which can be expressed as a single sentence", if the input is not already so. Under such assumptions, aggregation involves combining simple representations into complex ones.

A Brief Description of Aggregation

Aggregation has the surface form of combining components of a text at various levels, from words, phrases to clauses. Using the broad definition of aggregation (Reape and Mellish, 1999), "aggregation is the combination of two or more linguistic structures into a single linguistic structure which contributes to sentence structuring and construction".

A more specific description, revealing more of the effect of aggregation on the text as a whole rather than just at the sentence level, is as follows:

Definition 1.1 *Functioning as one or a set of processes acting on some intermediate representations in text structuring, aggregation decides which pieces of representation can be combined together to be realised as a single sentence later on so that a concise*

and cohesive text can be generated while the meaning of the text is kept almost the same as that without aggregation.

This definition makes use of the concepts of conciseness and cohesion. A concise text uses less words or sentences to express the same meaning. Cohesion is a semantic relation which connects a piece of text to what is mentioned before (Halliday and Hasan, 1976). In a cohesive text, pieces of texts are tied together nicely and there is no presupposition that cannot be resolved. Aggregation tends to make a text more concise and cohesive through three means, which at the surface level look like (some examples below are from (Dalianis, 1996)):

1. Using more general phrases to substitute for a set of words or phrases, and therefore making a text more concise. For example,

   John slapped Peter. Peter hit John back. => John and Peter fought.

   where the more general term *fight* is used to express the meaning of *slap* and *hit.*

2. Using connection words to join clauses into complex sentences and leaving out repeated parts. Here we mainly talk about Parataxis, which uses a coordinator like "and" or "or" to connect two clauses of equal status. For example,

   t1 has the phone number 100. t1 has a hot number 200. => t1 has the phone number 100 and a hot number 200.

   Parataxis contributes to both conciseness and cohesion, because it results in less words and sentences and the interpretation of the second coordinated part depends on what comes before it.

3. Using rank-shifting to embed one clause inside another. Rank-shifting/Embedding (Halliday, 1985) means that an item of a rank (i.e., sentence, some sort of sub-sentence, word, letter) higher than or at least equivalent to that of a nominal or verbal group comes to function as a modifier of the group. We show an example from (Scott and de Souza, 1990) in Figure 1.2. In the example, the first clause in the set is the nucleus. Aggregation converts the set of clauses into one sentence by either embedding or subordinating other clauses into the nucleus. In the

figure, arrows show how embedding maps clauses to NP components, including adjectives and noun phrases/appositive components.

Embedding makes a text more concise in terms of the number of sentences, and often also in terms of the number of words. But some embeddings do not delete words, for example, transforming a sentence into a relative clause of an NP, as in,

George, who is my friend, received a letter from Peter.

Embedding can also have an effect on the cohesion of a text because over-embedded NPs can destruct the cohesive ties between pieces of a text.

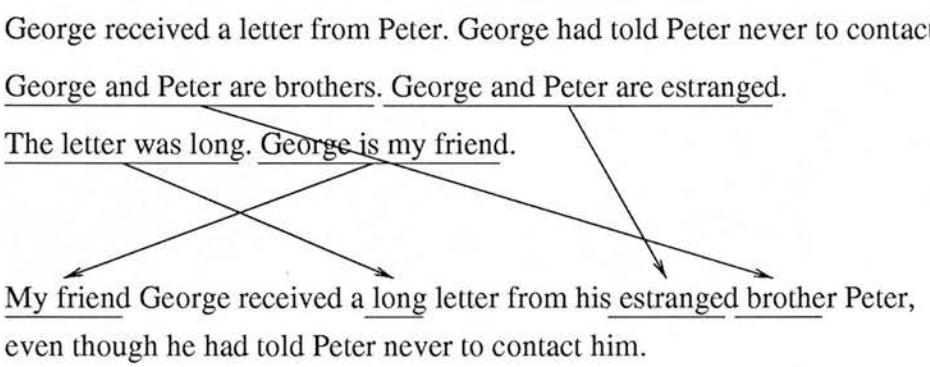


Figure 1.2: Examples of embedding

In the set of RST relations defined in (Mann and Thompson, 1987b), we pay special attention to two relations because of their connections with embedding and parataxis (this will be discussed in greater detail in the following chapters):

ELABORATION : a nucleus/satellite relation, where "the satellite presents additional detail about the situation or some element of subject matter which is presented in the nucleus or inferentially accessible in the nucleus". Mann and Thompson (1987b) identify six types of ELABORATION, one of which is called OBJECT-ATTRIBUTE ELABORATION, where the satellite presents an attribute of an object in the nucleus.

JOINT : a multi-nuclear connection between text spans where there is no other relation holding between them.

Why Aggregate in NLG?

Looking at student composition teaching, one finds that the skills of using subordination (including adjectives and adverbs, phrases of all types and subordinate clauses) and coordination to produce complex sentences are closely related fundamental principles of writing (Dillon, 1981), which are called aggregation techniques in NLG. This emphasises that aggregation is an essential writing technique rather than a peripheral facility, so a revision-based architecture which positions aggregation completely in a revision stage ((Robin, 1994b), more details in Section 2.3) might under-address the effect of aggregation on text production.

Indeed, in human written texts, aggregated sentences appear frequently. For example, Shaw and McKeown (1997) analysed a corpus of the first few sentences of the discharge summaries of 54 patients and found that there are on average 3.5 propositions per sentence and a maximum 12 propositions in a single sentence. "Conjunction is the most popular aggregation operation, followed by PPs, and then adjectives." There are also less frequent uses of participles and relative clauses in their corpus. In Chapter 4, we will describe a corpus analysis, which shows that human authors often use embedding in writing museum labels.

The purpose of writing a text is often to transfer information to other people. For a text to be easily understood by others, it has to be coherent. Using the description of Raskin and Weiser (1987), "Coherence refers to the consistency of purpose, voice, content, style, form, and so on of a discourse as intended by the writer, achieved in the text, and perceived by the reader." It is closely related to the notion of "readability" (Just and Carpenter, 1987), which refers to the difficulty or ease with which a text is read. Factors in both text and reader may contribute to readability. Such text properties include format, typography, content, style, vocabulary difficulty, sentence complexity and cohesiveness, etc. Reader factors include motivation, abilities, background knowledge and interests, etc. Readability can be estimated in terms of reading grade level, based on selected and quantified variables in text, especially some index of vocabulary difficulty and of sentence difficulty. In this thesis, we use "coherence" and "readability" as equal terms. Since aggregation plays a role in both textual and

stylistic appropriateness of a text, it contributes to the coherence of a text.

Psycholinguistic research has shown the effect of aggregation on generated text. (Scott and de Souza, 1990) discusses the factors that affect the effectiveness in getting information across and argues that the importance of conciseness cannot be over-stated. (Dalianis and Hovy, 1996) also mentions that given a set of randomly ordered propositions, in order to generate text human subjects would reorder them to facilitate the aggregation of propositions with some identical parts, e.g., subjects and verbs.

From the implementation perspective, an NLG system often needs to generate texts from general purpose knowledge bases, where the knowledge cannot be optimally organised to satisfy all the requirements of text generation. We have mentioned that generation often starts with simple units smaller than a sentence. If a text is directly generated from such units, there may be much redundancy in it, which may hinder the understanding of the text. Aggregation makes a text more concise and therefore helps to create a more readable text. Dalianis (1996) addresses the importance of aggregation as: "People and systems must perform aggregation to make their text more readable, understandable and fluid; not doing so risks the reader's misunderstanding or irritation."

Aggregation is also necessary for satisfying stylistic preferences. There are domains which demand as much information as possible to be got across in relatively little space, for example, newswire articles (Robin, 1994b), encyclopedia descriptions and museum labels. In such genres, aggregation is not only important to keeping redundancy to the minimum but also playing a big role in creating short but coherent descriptions which convey a large amount of information about domain objects.

In addition, aggregation also has an effect on such textual qualities as empathy and perceived friendliness, etc. In NLG, there is no clear model to address the purpose of aggregation and we do not intend to answer this question formally here.

In recent years, aggregation has attracted more and more attention due to its significant effect on improving the readability of generated text. It has become an almost indispensable component of all NLG systems, although the ways it is realised can be very different. For example, it can be an individual module or distributed across several

generation modules.

1.1.3 Problems with Aggregation Research

Although NLG has attracted substantial amount of interest in recent years, only a small amount of research focuses on aggregation (more details will be given in Chapter 2). Wilkinson (1995) first mentioned the conceptual difficulties of aggregation as an individual research topic in NLG. The problems include inadequate definition (aggregation does not equal redundancy elimination), vagueness as to its location in text generation (it can happen at various places) and unclear relationship to other phenomena (there is difficulty in isolating the aggregation process from others). Reape and Mellish (1999) developed this argument by reviewing the published research on aggregation. They presented two definitions and attempted to answer some important questions for aggregation. In this section, we talk about these arguments in detail because they partly motivate the work described in this thesis.

Problems of Defining Aggregation

The term "aggregation" was first used in (Mann and Moore, 1980), simply meaning putting clauses together. Since then, researchers keep on defining the term, but no definition is agreed by all.

A typical definition of aggregation is given in (Dalianis, 1996) where aggregation is "the process of removing redundant information in a text without, (ideally), losing any information". Hovy describes aggregation as a sentence-level planning task for compacting the communication materials – "aggregation uses the fact that information units, represented by the domain system as separate individuals, are often generated in the text as a group sharing pertinent features, and can therefore be abbreviated" (Hovy, 1993). So their definitions emphasise its effect of minimising redundancy.

Along the same line, (Robin and Favero, 2000) gives another definition: "grouping several content units, sharing various semantic features, inside a single linguistic structure, in such a way that the shared features are maximally factored out and minimally repeated in the generated text."

Yet aggregation and redundancy elimination are two overlapping phenomena, not equivalent. The contribution of aggregation to the conciseness and cohesion of a text is much more than just removing redundancy, whereas other phenomena like anaphora can also do abbreviation. Redundancy elimination is only a side effect of using aggregation skills.

Reape and Mellish (1999) summarise a "narrow" and a "broad" definition from the literature. The narrow definition says: "Aggregation is any process which maps one or more structures into another structure which gives rise to text which is more x-aggregated than would otherwise be the case". "X-aggregated text is text which contains no multiple nonpronominal overt realizations of any propositional content and no overt realizations of content readily inferable or recoverable from the reader's knowledge or the context which are not required to avoid referential ambiguity or to ensure grammaticality."

This narrow definition cannot cover some aggregation phenomena, in particular embedding as discussed in (Scott and de Souza, 1990). For example, the text "*George received a letter. It is long.*" satisfies all the conditions of an X-aggregated text, but it is not as concise as the one using embedding: *George received a long letter.* In the previous section, we mentioned the broad definition and also gave our definition there (Definition 1.1) because the broad definition does not reveal the textual effects aggregation intends to achieve. For concreteness, we will use Definition 1.1 for aggregation in this thesis, although it is still not perfect.

Other Problems Identified in the Literature

Hovy (1990) raises some important questions for aggregation: "What general rules of aggregation exist? How can the internal structure – symmetry, bushiness, etc. – of the paragraph structure tree be used to guide the application of such rules?" Later research tries to answer or elaborate these questions in various ways.

The five questions given in (Reape and Mellish, 1999) demonstrate the problems with research in aggregation very well because these are basic questions for any computational theory yet there are no generally agreed answers for them in the state of art of

aggregation. The questions are:

- "why is aggregation done": is about the definition and function of aggregation. In the literature, aggregation is often related to conciseness and cohesion, which are the aspects of a text that directly benefit from aggregation. Most existing definitions mention these two concepts and also coherence.

- "when/where is it done" and "what is it done to/on": are about the location of aggregation in text generation, and the resources and structures on which aggregation is performed.

  The answer proposed by Reape and Mellish (1999) is "whenever and wherever the appropriate structures arise". This does not seem to add much to the general understanding that there is not a single place where aggregation should happen. For example, Horacek (1992) and Dalianis and Hovy (1996) say that aggregation is appropriate whenever there is redundancy in the information to be presented, in particular when the information is explicitly repeated.

  For the "where" question, Dalianis and Hovy (1996) state that "in a simplified linearized model of the generation process, aggregation takes place after content determination (that is, after the content has been selected and preliminarily organized into a discourse structure) and before realization." This means that aggregation is a sub-task of sentence planning. Although this approach is followed by much research on aggregation, there are counter-arguments to it, e.g., (Bateman et al., 1998; Cheng and Mellish, 2000).

- "in what order are its subparts done": is about the effect of one type of aggregation on another. If a set of aggregation rules are used, then what is the order of applying these rules? As more rules are devised to cover more types of aggregation, this problem will become serious.

More Problems

In addition to those addressed above, we have identified the following questions:

1. *what is the relationship between aggregation and other generation processes*: this is related to the "when/where" questions and includes the relations among different subtypes of aggregation. Although (Wilkinson, 1995; Reape and Mellish, 1999) mention the need to clarify these questions, the answers they give are not clear in themselves and cannot provide satisfactory explanation.

   In a pipeline architecture, aggregation is a part of sentence planning, which cannot normally affect the decisions made in content determination. No previous research tries to identify which processes are most likely to be affected by aggregation or explores how aggregation interacts with major generation tasks like text structuring and how to make use of these interactions for better NLG.

2. *how reliable are the aggregation rules*: most aggregation operations are based on rules, which are usually subjective observations of researchers from a domain corpus. The replicability of such rules are questionable, but little previous research touches this problem.

3. *what effect does aggregation have on the coherence of generated text*: no serious evaluation has been done on how aggregation affects the coherence of generated text. This is due to the great difficulty in judging the coherence of a text in general, therefore evaluation tends to focus on conciseness.

This thesis is motivated by the above questions/problems for aggregation research and it endeavours to answer some of these questions and move toward a better understanding of aggregation as a complex generation issue.

1.2 Embedding – the Theme of This Thesis

Several types of aggregation have been identified in the literature, but it is impossible to work on all of them in detail. To study the problems associated with aggregation, we chose to focus on a specific type of aggregation – embedding. The reasons are:

- There is less work on embedding in the literature, which usually focuses on parataxis. For embedding, the answers to the questions identified by Reape and Mellish are far from clear.

- We have found that embedding interacts with other generation tasks, e.g., document structuring and referring expression generation, in a complex way (as will be discussed through the following chapters). So embedding provides a rich source for studying the interactions between generation tasks.

- Since embedding is a subtype of aggregation, we hope that a study on embedding can shed light on the problem of aggregation in general, for example, on the relationship between aggregation and other generation processes.

Within the boundary of embedding, this thesis is mainly concerned with modification decisions in constructing complex NPs. Linguistic studies usually classify an NP component apart from the head and the determiner as a restrictive component or a non-restrictive component. A restrictive component is for uniquely identifying the entity denoted by the NP head, whereas a non-restrictive component is any additional information that is given to a head that has already been viewed as unique or as a member of a class that has been independently identified, and therefore is not essential for the identification of the head (Quirk et al., 1985). In NLG, the decisions about non-restrictive components, including selecting which information is to be embedded and deciding how to realise the information, are the concerns of embedding (more details in Chapter 3).

(Halliday and Hasan, 1976; Halliday, 1985) particularly distinguish embedding from hypotaxis where "a clause is dependent on another clause but not structurally integrated into it; it is not a constituent of it". Halliday (1985) says:

> "Embedding is a mechanism whereby a clause or phrase comes to function as a constituent within the structure of a group, which itself is a constituent of a clause. Hence there is no direct relationship between an embedded clause and the clause within which it is embedded. The relationship of an embedded clause to the 'outer' clause is with a group as an intermediary, that is, the embedded clause functions in the structure of the group and the group functions in the structure of the clause. So embedding is not a relation between two clauses."

According to this, adjectives and prepositional phrases are embedded or rank-shifted, but non-restrictive relative clauses are hypotactic.

From the generation point of view, this distinction is not necessary. Since we assume that the input to document structuring is clause-sized semantic representations, embedding is an effective way of combining different properties of an object usually represented as simple separate facts, and expressing them more concisely as NP modifiers like relative clauses or adjectives in the main clause containing the object. This has the surface form of clause combining. Following (Scott and de Souza, 1990), we do not distinguish between NP components and treat all of them as embedded.

Embedding, or aggregation in general, is achieved by combining the identical parts of several representations and producing a single representation with more complex inner structure. For example, if the input to aggregation is an RST tree from text structuring and the leaf nodes of the tree are individual facts, aggregation combines the adjacent leaf nodes with similar inner structures and outputs a revised tree with less branches but more complex structures in some leaf nodes. This operation has to take into account discourse, semantic and syntactic restrictions, so it is a complex decision.

1.3 The Domain: Object Descriptive Texts

We are interested in the domain of object descriptive text, which is a discourse intended to give a mental image of an object or a group of related objects through providing relevant information about them. Descriptive texts are prevalent in the world as they include almost all texts that are supposed to provide information about specific domain objects. Some obvious examples are museum item descriptions, encyclopedia articles, animal or plant descriptions in a zoo or botanical garden and shop catalogues, etc. These texts aim at providing information, although rhetorical effects such as persuading can be achieved at the same time.

This is a potentially interesting domain for NLG because descriptions written for a given purpose and group of readers tend to display a pattern. Compared with other domains like argumentative texts, there are more repetitions in the description domain

in both the types of information conveyed and the syntactic structures used. So we can disregard the intentions behind the descriptions and simplify the generation problem by only modelling the patterns in a given corpus.

Beside the wide coverage of the domain, our choice is also driven by some desirable properties of object descriptive texts. Such texts are usually rich in aggregation phenomena, especially subordinated NP components, which gives us a rich space to explore. In the mean time, they are simple in rhetorical phenomena, in terms of the number and types of rhetorical relations used. Therefore, they present a good starting point for studying the interactions between aggregation and other generation tasks, without getting into the full complexity of rhetorical planning.

In this thesis, we restrict our discussion to museum descriptive texts, which describe items displayed in museums. We expect these to carry typical features of object descriptive texts in general. Below is an example of a museum description from the IvyWu gallery of the National Museum of Scotland.

> Throne and Cover
>
> Small portable thrones were used in the private apartments of the Imperial Palaces.
>
> This example from the time of the Qianlong Emperor 1736-95, is made of lacquered wood with decoration in gold and red. The design on the seat is a five clawed imperial dragon in a circular medallion. On the inside of the arm pieces are small shelves on which precious possessions can be placed and studied as an aid to contemplation.
>
> The throne cover, from the reign of Jiaquing, 1796-1820, is woven in yellow silk which is the imperial colour of the Qing Dynasty, 1644-1911. It would have covered the throne when not in use.

We will sometimes come across examples produced by an NLG system called ILEX (Intelligent Labelling EXplorer) (Oberlander et al., 1998) in our discussion. ILEX is an adaptive hypertext generation system, providing natural language descriptions for museum objects. It was developed at the former Department of Artificial Intelligence and Human Communication Research Centre of the University of Edinburgh. More details of this system will be given in Section 7.3.1.

1.4 Contributions

This thesis attempts to somehow fill in the research gap identified in Section 1.1.3. It targets one aspect of aggregation that has not been explored by previous research: the complex interactions between aggregation and other text generation tasks, and what these interactions imply for a generation architecture. To study these problems in detail, the thesis focuses on embedding. It identifies regularities in the way human authors produce complex NPs using embedding and reveals the interactions between embedding and such processes as document structuring and referring expression generation. These findings motivate a different generation architecture from the pipeline architecture defined in (Reiter, 1994) in order to capture the interactions in a better way. Issues closely related to this main thread are also studied, for example, the replicability of the observations about embedding regularity and the evaluation of the effects of embedding on the coherence of generated text.

The contributions of this thesis are mainly on seven aspects, along the line of observing regularities for embedding → clarifying the interactions between embedding and other processes → extracting preferences → implementing the preferences → evaluation.

- We classify NP modifier uses into three types, which are the concerns of different generation processes. Through corpus annotation, we find that human subjects can relatively reliably tell the different modifier uses. Although previous research has mentioned that different functions are served by NP components, there was previously no evidence that such a distinction can be identified reliably by humans.

- We perform a corpus analysis, which enables us to summarise general embedding rules for the content selection and realisation of NP modifiers in descriptive text generation. We adopt a more systematic approach to corpus analysis than just using our intuitions, therefore our rules are more reliable than those obtained from previous research.

- We investigate the complex interactions between

  - embedding and referring expression generation, and attempt to capture this

interaction in generation algorithms;

– embedding and planning entity-based and relation-based coherence.

These interactions demand embedding to be considered earlier in generation than it currently is, i.e., in content determination rather than just in sentence planning, in order to generate more coherent descriptive text. This is compatible with what other researchers have pointed out, e.g., (Horacek, 1990; Horacek, 1992; Reape and Mellish, 1999). However, we do not stop here, but instead propose approaches to implement this.

- We study the phenomenon of using non-restrictive NP components to express semantic relations, which has not been discussed in the literature. We identify the significant factors and summarise heuristics that can be used for generation.

- We propose preferences among coherence features as a way of capturing the complex interactions we have discovered between generation tasks, which mainly include features of entity-based and relation-based coherence and embedding.

- We implement the preferences in two different generation systems and compare the behaviours of a pipeline and a non-pipeline architecture. This demonstrates that it is possible to capture the preferences in a non-sequential way, which will lead to coherent text.

- We quantitatively evaluate the observed embedding rules using the annotated corpus. What is more, we evaluate the coherence of generated text with embedding using human subjects. We also make an attempt to automatically evaluate the readability of a text.

To summarise, through this thesis, we wish to give a clearer account of embedding, reveal its connections with other generation tasks and find better ways to model these connections. However, not all of the theoretical issues discussed in this thesis can be exploited in the actual systems because of the limitations in system architecture and input.

1.5 Organisation of the Thesis

The central ideas of this thesis are developed through Chapters 3 to 8. Chapters 1 and 2 motivate our work, and Chapter 9 concludes the discussion.

Chapter 2 gives an overview of research on aggregation, which supports the important points discussed in the current chapter, i.e., the problems and the contributions. We introduce the literature according to the method used by each work, e.g., a rule-based or similarity-based approach. We argue that the interactions between aggregation and other generation processes are given little attention in previous work, so they will be the topic of this thesis.

Chapter 3 divides a referring expression into a referring and a non-referring part and discusses the complex interactions between the two parts. These motivate a set of syntactic and semantic constraints on the generation of the non-referring part.

Chapter 4 describes two corpus analyses we performed to discover the regularities in the usage of NP modifiers in museum descriptions. The first analysis reveals the general characteristics of NP modifiers in such texts and the summarised embedding rules are used in the two implementations described in Chapter 7. The second analysis uses more systematic and fine-grained approaches and therefore provides reliable evidence for supporting the discovered embedding rules and modifier generation algorithms. However, the results are not used in our implementations.

Chapter 5 focuses on the phenomenon of using non-referring NP components to support the situation presented in the main proposition containing the NP, in particular, the acceptability of using non-referring NP components to express semantic relations that might normally be signalled by *because* and *then* between separate clauses. It describes a psycholinguistic experiment regarding the similarity between the meanings expressed through two different syntactic constructions. The experiment tests several relevant factors and enables us to accept or reject a number of hypotheses. This study focuses on the theoretical aspect and is not taken up in later chapters.

Chapter 6 describes the interactions between aggregation and document structuring, and argues that resolving the complex interactions within and between tasks is more

important to the generation of a coherent text than modelling each individual factor. It captures the interactions discussed through Chapters 3, 5 and 6 as preferences among features considered by different generation tasks. Heuristics for the preferences are derived from general linguistic and discourse theories.

Chapter 7 describes the implementation of the above preferences in two generation systems: ILEX-TS (using a pipeline architecture) and GA-plan (using a search-based architecture). We start with a brief introduction to text planning architectures, which feature the major differences between the two systems, and then describe each specific implementation in more detail. We show through comparison that the non-pipeline architecture captures the interactions between tasks better.

Chapter 8 describes the evaluation of the multi-sentential texts generated taking into account the interactions between embedding and document structuring, i.e., the output of GA-plan. Using both automatic evaluation and human judgement, we show that the preferences indeed capture some truth about the notion of a coherent text. We also briefly compare our work with related work.

Finally, Chapter 9 readdresses the contributions of the thesis and suggests some possible extensions to the current work.

Chapter 2

About Aggregation: Taxonomy and Literature

This chapter gives an overview of research on aggregation, which motivates the problems discussed in Chapter 1 and enhances the contributions of the thesis. We first present a taxonomy of aggregation and then introduce the literature according to the method used by each piece of work, e.g., a rule-based or similarity-based approach. How each piece of work addresses the problems in Section 1.1.3 is given special attention. We argue that the interactions between aggregation and other generation processes are given little attention in previous work, yet they are important to the production of text which is concise and rhetorically coherent. This motivates our work described in this thesis.

2.1 What to Review

Except for (Scott and de Souza, 1990; Robin, 1994b), previous work on aggregation mainly focuses on parataxis and only mentions embedding briefly. The existing research on embedding shares the problems of aggregation in general. Also as mentioned in the previous chapter, we are interested in the interactions between aggregation and other generation phenomena and those among subtypes of aggregation, e.g., embedding and parataxis. Therefore, we do not only look at research on embedding, but rather review work on aggregation as a whole to gain a more general picture.

This chapter describes the existing work which complies with our definition of aggregation in Section 1.1.3. Using terms such as *clause combining, redundancy removing* and *grouping* (Horacek, 1992), the work addresses one or another aspect of aggregation. We mainly look at the following aspects of previous work for their relevance to the problems we have identified and we will study embedding from these perspectives (we leave the problem of evaluation to Chapter 8):

- Which types of aggregation are modelled and how, especially, is embedding handled;
- At which stages of generation is aggregation performed and what is the input to aggregation;
- How are interactions between aggregation and other generation tasks modelled and what is the order among subtypes of aggregation if more than one subtype is handled;
- How are the rules devised and what is the order of applying the rules if a rule-based method is used.

But we cannot address the first problem before we have a classification of aggregation types. So we start with an introduction to existing taxonomies of aggregation.

2.2 A Taxonomy of Aggregation

In Section 1.1.2, we briefly described a few subtypes of aggregation using examples. In this section, we introduce two representative taxonomies given in the literature and then the one we will use in this thesis.

Dalianis (1996) distinguishes four types of aggregation, based on the ways that redundancy is removed:

Syntactic aggregation : to remove redundant information, but leave (at least) one item in the text to carry the meaning explicitly. For example, *John is a subscriber. Mary is a subscriber. => John and Mary are subscribers.*

This is similar to the *structurally motivated purely propositional grouping* of (Horacek, 1992), which produces a summarising statement, followed by an enumeration. For example, *John likes vision, John likes robotics, John likes theorem proving. => John likes the following subjects: vision, robotics and theorem proving.*

We doubt that such aggregation can be carried out at a purely syntactic level although its effect can be demonstrated syntactically. Often the semantic similarity between the propositions plays the major role in the combination. So *syntactic aggregation* might actually not be a good name for such aggregation.

Elision : to remove information that can be inferred and leave no explicit carrier in the text. The information remains there implicitly. This does not seem to include the type of ellipsis which leaves a carrier, e.g., an VP ellipsis. An example from (McDonald and Busa, 1994) reads:

I would really like to have you guys over for dinner, so let me know whether for you it is better before (you leave for) <Florida> or <it is better> after (you come back from) Florida.

In this sentence, (...) stands for elided and <...> for aggregated, both of which do not appear in the final sentence. The difference between *elision* and *syntactic aggregation* is that nothing remains from the former operation (the information is probably given in the previous context) whereas an explicit item is left from the latter.

Lexical aggregation : to replace a set of lexemes with one, while more or less keeping the overall meaning. Two subtypes are identified:

bounded lexical aggregation, where the set of lexemes is a subset of a closed set of concepts, e.g., *John uses his mobile phone on Mondays, Tuesdays, Wednesdays and Thursdays. => John uses his mobile phone on all weekdays except Fridays.* This subtype preserves the original meaning.

unbounded lexical aggregation, where the set is a subset of an open set of concepts so there may be accuracy lost in this operation, e.g.,

John hits Peter. Peter hits John back. => *John and Peter fight.*

This subtype somehow changes the original meaning.

Referential aggregation : to use techniques such as pronominalisation to remove redundancy. For example, *John and Mary are subscribers. John and Mary are idle.* => *John and Mary are subscribers. They are idle.*

Since the above classification focuses on redundancy elimination, it only partially overlaps with our definition of aggregation. It includes some phenomena that do not combine representations, for example, *referential aggregation* addresses an issue normally considered by the referring expression generation task rather than aggregation, and excludes some phenomena which do not remove redundancy, e.g., embedding using relative clauses.

Wilkinson (1995) and Reape and Mellish (1999) summarise a more fine-grained taxonomy from the literature, based on the level of representation on which aggregation operates, e.g., semantic expressions or lexemes. The typology is shown in Table 2.1, where the columns give the aggregation types, the representations they manipulate and their locations in the generation process.

Aggregation Type	Representation	Location
Conceptual	domain concept	before text structuring
Semantic	semantic entity	before or after text structuring
Discourse	discourse tree	between text structuring and sentence planning
Syntactic	syntactic tree	in sentence planning
Lexical	lexeme	in lexicalisation
Referential	domain entity	in referring expression generation

Table 2.1: Aggregation types summarised in (Wilkinson, 1995) and (Reape and Mellish, 1999)

Although this classification has the widest coverage so far, we find that the distinctions between types are not all that clear. For example, most previous research is classified as *syntactic* aggregation, especially the work that uses rules, but as we will introduce in Section 2.3, in fact much research applies rules on semantic representations at the bottom level of RST trees from text structuring, e.g., (Dalianis and Hovy, 1996). Although the rules are presented from the surface perspective, they manipulate representations

combining semantic and syntactic information. There is considerable overlap between *discourse, semantic* and *syntactic* aggregation as they are defined and realised at the current stage.

Another example is *lexical* aggregation. This is a two step process, firstly the relevant concepts are put next to each other or grouped together and then an appropriate lexical item is chosen to express the set of concepts more concisely. The first step is similar to *semantic* or *syntactic* aggregation, whereas the second is a part of lexicalisation. So *lexical* aggregation can be realised through the combination of other processes. It is probably not sensible to take *lexical* aggregation as a single process in lexicalisation. There is a potential danger to change the decision of text structuring by combining representations at this stage because such an operation might result in serious destruction to the satisfaction of some goals. We doubt that such aggregation should be carried out in generation.

Another deficiency of this classification is that the effect of one type of aggregation may be achieved by another using the same system resources. For example, Shaw and McKeown (2000) implement the *conceptual* aggregation described in (Horacek, 1992) in sentence planning (in the same way as semantic or syntactic aggregation) rather than before text structuring by accessing a domain ontology. As long as the resources are available, the inference required for conceptual aggregation can be carried out at any stage. This means that the mapping from aggregation type to linguistic representation or locus in generation is not one to one, but many to many, therefore a classification based on representation or locus is not well motivated.

Finally, as stated in (Wilkinson, 1995), aggregation can be decomposed into two subtasks: deciding *what to aggregate* and deciding *how to aggregate.* The *what* question often concerns higher level operations than pure syntactic rearrangement. For this reason, little aggregation can be carried out at a purely syntactic level – often the underlying semantic similarities are needed to limit the space of possible combinations.

The above discussion also explains why we do not give examples for the classification. The available examples all more or less cross types and therefore cannot demonstrate the distinctions.

Because of these deficiencies, we do not follow their classification, but rather emphasise the combining aspect of aggregation and use the more traditional distinctions made in (Halliday and Hasan, 1976; Halliday, 1985; Scott and de Souza, 1990) for clause-combining. This classification is based on the surface appearance of aggregation, but we do not assume that aggregation works at the surface level necessarily. In other words, we assume that aggregation can work on any intermediate representation in the generation process, and every aggregation decision corresponds to a clause-combining operation as in Figure 2.1. In this thesis, we will often mention aggregation as if it applies directly to the surface level, purely as a useful shorthand.

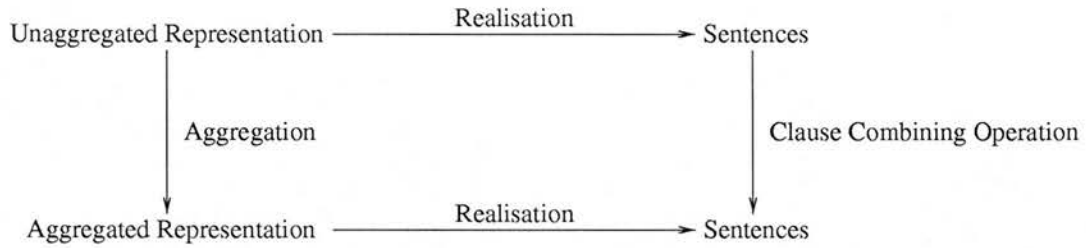


Figure 2.1: Mapping between an aggregation and a clause-combining operation

The classification uses the rank (e.g., sentence, phrase, word) and the degree of syntactic dependency of the combined parts as the measurements (as shown in Table 2.2). Although all aggregation has the surface form of clause-combining, different ranks and syntactic dependencies exist among the combined parts. Using the two measurements, domain independent aggregation is grouped into three types: *embedding, hypotaxis* and *parataxis.*

Dimensions		**Syntactic Dependency**	
		Dependent	Independent
Rank	Phrase	Embedding	Semantic parataxis
	Clause	Hypotaxis	Parataxis

Table 2.2: Types of aggregation

Embedding is to use one clause to function within the structure of a group in another clause, that is, the clause is rank-shifted to a phrase component of the other

clause. The embedded part can be a premodifier or postmodifier of the group or the head. For example,

The house is beautiful. The house is by the bridge. => The beautiful house is by the bridge. (NP premodifier)

The man came to dinner. The man stayed for a month. => The man, who came to dinner, stayed for a month. (NP postmodifier)

He went to the church. He was sooner than the rest of us. => He went to the church sooner than the rest of us. (adverbial phrase postmodifier)

Hypotactic aggregation (Hypotaxis) is to bind clauses of unequal status in which one is dominant and the other is dependent. Some examples from (Halliday, 1985) are:

The horse stopped. He fell off in front. (condition) => *When the horse stopped, he fell off in front.*

I didn't revise my notes for the exam. I lay down and went to sleep. (antithesis) => *Instead of revising my notes for the exam, I lay down and went to sleep.*

In the above sentences, the status of the two clauses are unequal. The main clauses can exist without the subordinate clauses, whereas the subordinate clauses cannot exist alone. Possible cue phrases are chosen to make explicit the rhetorical relations between two clauses. In some cases, the less important clause can be missed out completely if its meaning is redundant.

Paratactic aggregation (Parataxis) is to bind clauses of equal status. We distinguish two types of parataxis:

Semantic parataxis : using the abstraction of RST, this concerns clauses related by explicit multi-nuclear relations (e.g., SEQUENCE and CONTRAST) or by implicit connections like parallel common parts. In the result sentence, some identical parts are left out and the distinct parts are combined to complex phrases, for example, *John is a subscriber. Mary is a subscriber.* => *John and Mary are subscribers.* Here the two sentences have parallel common parts and therefore aggregated.

Although it is possible to aggregate propositions with the above connections at a syntactic level, we use the term *semantic* to distinguish between cases where there are at least some semantic relations between propositions and those without obvious connections, as described below.

Textual parataxis : this is a purely textual combination of adjacent sentences, probably connected by a JOINT relation, using coordinators. The initiating and the continuing clause are independent, for example, *This necklace is in the Arts-and-Crafts style and was designed by Jessie King.*

A similar division is adopted in (Shaw and McKeown, 1997; Shaw, 1998a), where separate submodules are designed to perform paratactic and hypotactic aggregation. The advantages of the above classification are that the distinctions between types are clear and each type can be studied as an individual phenomenon. One disadvantage is that its coverage becomes too wide, in particular, *hypotaxis* stretches to the phenomenon of combining clauses using rhetorical relations, which is the major concern of text structuring. In addition, using surface appearance may not help to reveal the underlying processes or the underlying similarities between types.

The three types of aggregation can happen at various stages of generation, and each is a complex decision under the constraints of semantics, syntax and pragmatics. Using an example from (Scott and de Souza, 1990), we can see that lexical semantics must be considered when choosing an optimal embedding type.

A man bought the picture. He has blond hair.
can be aggregated as: *A blond man bought the picture.*
If the second clause changes to: *He has black hair.*
we cannot say: *A black man bought the picture.*
but can only say: *A man with black hair bought the picture.*

Linguistic research on discourse and sentence structure has formed the foundation of recent aggregation research. For example, Matthiessen and Thompson (1987) study clause combining from the computational linguistic perspective. Following Halliday (1985), they mention two degrees of clause combining as parataxis and hypotaxis. Parataxis

includes coordination, apposition and quoting. Hypotaxis includes clause combining involving non-restrictive relative clauses, clauses of reported speech and enhancing hypotaxis (hypotactic clause combining involving some kinds of circumstantial relations). Instead of studying the whole set of clause combining possibilities, they place an emphasis on enhancing hypotaxis and study the relation between the organisation of hypotactic clause combining and of discourse in general. Based on their observation of rhetorical organisation, they make the hypothesis: "Clause combining in grammar has evolved as a grammaticalization of the rhetorical units in discourse defined by rhetorical relations." This partially explains why much recent aggregation research bases its aggregation algorithms on some kind of rhetorical organisation of discourse.

In this thesis, we mainly refer to *embedding* and *parataxis* when we mention aggregation. In the following sections, we introduce previous research on aggregation. The introduction is organised according to the way aggregation is approached, e.g., using a rule-based or similarity-based method. Table 2.3 illustrates the other dimensions for our survey. That is, in addition to the difference in the way that aggregation is approached, we also focus on such aspects of an existing work as the types of aggregation it models, the representation aggregation manipulates, the source of aggregation rules, the architecture of the implemented system and whether or not interactions between aggregation and other generation tasks are studied.

We give details of the aggregation modules of PLANDoc, MAGIC, STREAK and PROVERB as aggregation plays an important role there. When possible, we mention other NLG systems using similar aggregation techniques, but we do not intend to give a complete survey of the aggregation modules of all implemented NLG systems.

2.3 Rule-Based Aggregation

In this section, we introduce work using rules or heuristics for aggregation. We identify three ways of devising rules:

Linguistic observation : rules are given by linguists based on their knowledge of languages in general. These rules are generally applicable (i.e., cross-domain or even cross-language), but may not be enough when applied to a specific domain.

Systems	Dimensions for Survey				
	Aggregation types	*Representations*	*Sources of rules*	*Architectures*	*Interactions*
Dalianis'	embedding, parataxis	predicate-argument structures connected by RST relations	observing human texts	pipeline	rule ordering
PROVERB	embedding, parataxis	structured preverbal messages	adapting and extending Dalianis' rules	pipeline, aggregation in microplanning	none
HealthDoc	semantic grouping, sentence structuring and delimitation	sentence plans	using the rules of Dalianis and Scott	parallel planning using a blackboard	among sentence planning subtasks
STREAK	embedding, parataxis	a three-layered representation	analysing the lead sentences of newswire reports	revision-based, combining representations in the revising stage	none
PLANDoc and MAGIC	semantic, hypotactic and paratactic aggregation	a list of propositions in predicate-argument structures	analysing a corpus of patient information	pipeline, aggregation in microplanning	the ordering of subtypes of aggregation
HYSSOP	parataxis	deep semantic representations	no rules	pipeline, aggregation in sentence planning	none

Table 2.3: Dimensions for literature survey

Psycholinguistic influence : heuristics are obtained from controlled experiments with human subjects. These rules shed light on how humans approach a language problem or how a phenomenon affects human cognition. Such heuristics have been used in NLG systems.

Corpus analysis : rules are summarised from recurring patterns in some sample texts of a specific domain. These rules are usually more fine-grained, but some are not portable across domains.

Each method above has its pros and cons and there are usually overlaps between the rule sets discovered by different methods. We organise research on rule-based aggregation into these three categories and describe it below.

2.3.1 Linguistic Observation

In the Fragment-and-Compose paradigm of (Mann and Moore, 1981) for generating English text, aggregation plays a significant role. The composition operation is partially based on a set of aggregation rules corresponding roughly to the clause-combining rules of English. A revised Hill Climber algorithm based on preference rules is used to select the best application of aggregation rules.

The aggregation rules cover some specific proposition patterns in English, mainly clause coordination. For example, the *conjoin mid-state* rule combines sentence pattern *Whenever X then Y. Whenever Y then Z.* to *Whenever X then Y and then Z.* This rule-based approach is still followed by much contemporary aggregation work.

To cover the clause combining phenomena in an application domain, implemented NLG systems often use a subset of the general linguistic rules for clause combining. The subset contains the frequently recurring patterns in the targeted domain. For example, the *grouping rules* of (Dalianis and Hovy, 1996) and some revision rules of (Robin and McKeown, 1993), e.g., the *conjoin* and *adjoin* revisions, (to be introduced in Section 2.3.3) are such subsets.

2.3.2 Psycholinguistic Influence

Influenced by psycholinguistic research and the psychology of memory, (Scott and de Souza, 1990) gives thirteen heuristics for both the necessity of putting clauses together and doing embedding and paratactic coordination on hierarchical rhetorical structures. These heuristics have had a great effect on later aggregation research and some of them are widely followed in aggregation algorithms.

The heuristics that are most relevant to embedding include:

- Embedding is restricted to clauses connected by the ELABORATION relation.
- When embedding, the nucleus of ELABORATION must form the matrix of the sentence and the satellite the embedded part.
- In realising embedded parts, syntactically simpler expressions are to be preferred over more complex ones. The preference order is: lexicalised $\prec$ phrasal $\prec$ clausal ($\prec$ reads "precedes"). This is particularly useful for selecting among multiple syntactic realisations of the embedded information and for devising embedding rules to achieve a relatively optimal result.
- "Propositions of a LIST relation should not be embedded if doing so would make the number of remaining propositions in the relation equal to 1." That is, there should not be dangling sentences after embedding.
- "Self-embedding is only allowed in cases where the proposition that is the deeper of the two embeddings is expressed as an adjective or adverb." That is, only one level of clause embedding is allowed, although in practice, people are able to produce and understand phrases with multi-level clause embedding.

We also make use of some of these heuristics in our embedding algorithm, so we will revisit them in later chapters.

This is mainly theoretical research about clause combining, and no implementation is intended. The pieces of work introduced below are different because they are application oriented and are all implemented. We will describe how they handle some

frequently asked questions for a rule-based method, that is, on which representation the rules operate and in which order the rules are applied.

2.3.3 Corpus Analysis

One can use just general clause combining rules for any application domain, but this might be neither efficient nor sufficient and corpus analysis is often needed. There are two reasons. Firstly it is not efficient to have rules that are never used in a domain because they slow down the generation process. So general rule sets need to be tailored to suit a specific domain. Secondly the dramatic differences between domains make general rules insufficient for handling complex domain specific phenomena. For example, the clause combining cases in newswire articles are very complex and need many specific rules. Therefore, corpus analysis is normally needed for designing NLG systems targeting specific domains.

In order to find out how humans express information concisely, Dalianis and Hovy (1996) asked humans to create texts out of sets of scrambled propositions. They then manually built RST trees for these human texts and also for the original random proposition sets. The differences between the RST trees of the human texts and the random texts motivate four groups of aggregation rules.

1. Grouping Rules: for collecting and collapsing clauses with common elements, sometimes using cue words to signal aggregation. These rules consist of a *subject grouping* rule and a *predicate grouping* rule which aggregate propositions with identical subjects or predicates to form a single proposition. Below is an example of applying the *subject grouping* rule.

   t1 is an idle subscriber. t1 has a phone number 100. => t1 is an idle subscriber with a phone number 100.

   These rules result in both subject/predicate coordination and embedding. They operate on the JOIN (similar to the JOINT relation of RST) and ELABORATION relations, and have been addressed by others as *grouping motivated by structural reasons* (Horacek, 1992), *embedding* (Scott and de Souza, 1990) and *forward and backward conjunction reduction* (Kempen, 1991).

2. Ordering Rules: for changing the order between clauses or between phrases within clauses to satisfy such priorities as supertype $\prec$ attribute (reads "supertype precedes attribute", e.g., *sp1 is a speech connection and is idle.*) and animate $\prec$ inanimate (e.g., *t1 is a subscriber. sp1 is a speech connection.*) These rules operate on the JOIN/LIST and ELABORATION relations.

3. Casting Rules: for preferring the same syntactic constructions and lexical items to be used for semantically similar items throughout the whole discourse. These are actually rules for style.

4. Parsimony Rules: for preferring the casting of simpler syntactic constructions. For example, *A subscriber t1 has a phone number 100* is better than *t1 is a subscriber and has a phone number 100.*

Some of these are domain independent rules, such as *grouping rules* and *parsimony rules*, whereas others such as *casting rules* might not be. These rules capture some simple types of embedding and parataxis. They work on a list of facts in predicate-argument structures, connected mainly by the JOIN/LIST and ELABORATION relations. Aggregation using these rules results in more complex argument structures, for example, an embedded fact as an argument.

The possibilities of applying rules in different orders to produce different texts are mentioned in (Dalianis and Hovy, 1996; Dalianis, 1997b), where it is called the *rule ordering problem* of aggregation. Dalianis and Hovy (1996) give partial orders of applying the rules as (A $\prec$ B stands for "A is applied precede B"):

> Repetition rules (a type of Parsimony Rules which omits repeated propositions completely) $\prec$ Grouping and Ordering rules $\prec$ Casting rules
>
> Repetition rules in content selection

The application of rules generally follows the *grouping, ordering* and *casting* order.

In (Dalianis, 1997a), a small-scale experiment aimed at finding out the optimal ordering of rules is described. The proposed order is: *unbounded lexical aggregation* $\prec$

predicate and direct object grouping $\prec$ *subject and predicate grouping* $\prec$ *bounded lexical aggregation.* However, the reliability of such an ordering is not validated.

Dalianis' work is the first piece of work focusing on aggregation. It studies some important issues, such as the classification of aggregation, general rules that can be used for aggregation, rule ordering and the use of cue words to avoid ambiguity. Although no definite solution is given to any of these issues, his work pinpoints a starting point and directions for later aggregation research.

(Fiedler and Huang, 1995; Huang and Fiedler, 1996) describe the aggregation and paraphrasing strategies of the PROVERB system, which verbalises mathematical proofs. They use three sets of rules: *semantic grouping, semantic embedding* and *pattern-based optimisation* rules, and there are overlaps between these rules and the *grouping rules* of (Dalianis and Hovy, 1996). For example, one predicate grouping rule looks like:

$$\frac{P[a] + P[b]}{P[a + b]} \tag{2.1}$$

where the form above the bar is the text structure before aggregation, and the one below is that after. 'P' stands for a logical predicate, and '+' can be either a logical $\vee$ or $\wedge$. This rule can be illustrated by the following example (although the use of $\wedge$ does not seem to be conventional):

$Set(F) \wedge Set(G)$ *(F is a set. G is a set.)* $=>$ $Set(F \wedge G)$ *(F and G are sets.)*

However, the sources of their rules are not clear, and the rules seem to be application oriented.

Working on an ordered sequence of *preverbal messages* (PMs) from macroplanning, the microplanner of PROVERB progressively maps application program concepts in PMs into Upper Model objects and then into text structures similar to Meteer's Text Structure ((Bateman et al., 1990; Meteer, 1992), to be introduced in Sections 3.2.2 and 7.2). Aggregation is a part of microplanning and operates on structured PMs before they are mapped into Upper Model objects. Although aggregation does not directly operate on linguistic structures, the combination of two adjacent application objects implies the combination of their linguistic resources. Aggregation makes use of the three sets of rules to remove redundancies in the resulting text structure, although the

rules actually work at the semantic level.

This work is not specifically about aggregation, but rather about the application of NLG techniques to a non-traditional application domain.

(Wilkinson, 1995) describes the sentence planner of the HealthDoc system, which uses the aggregation rules of (Dalianis, 1996) and the heuristics of (Scott and de Souza, 1990) to remove redundancy. HealthDoc provides customised hospital patient information materials. The sentence planner is the central component of the system. It uses sentence plans represented in SPL (a Sentence Plan Language developed for the Penman project (Kasper and Whitney, 1989)) as the input and handles various issues including semantic grouping, sentence structuring, lexical choice and reference choice. These modules run in parallel and are coordinated by an administrator module through a set of blackboards using stylistic reasoning.

However, the main point of (Wilkinson, 1995) is to identify the conceptual problems of aggregation as an individual research topic (discussed in Section 1.1.3). Wilkinson suggests that aggregation should be viewed instead as a characteristic of a proper text resulting from careful sentence planning, including semantic grouping, sentence structuring and sentence delimitation. This idea is realised in HealthDoc, which uses a "generation by selection and repair" strategy involving selecting sentence plans from a "master document" and then applying sentence planning to restore the coherence lost during the selection.

The main contribution of this work is that it pinpoints the problems with current aggregation research and proposes a way to overcome the identified conceptual difficulties. In addition, the sentence planning strategy using a blackboard gives a way to capture the interactions between different sentence planning modules. However, the description is at a fairly abstract level and there is no comment as to how well this new strategy performs.

Robin analysed the lead sentences of newswire reports of basketball games to motivate a revision-based generation architecture as well as a set of revision tools (i.e., rule base) for automatic generation of the complex lead sentences that are normally used in sports reports (Robin and McKeown, 1993; Robin, 1994a; Robin, 1994b; Robin and

McKeown, 1996). Robin argues that complex sentences can be generated in two steps, first to produce a basic sentence expressing essential information and then to incrementally revise it to incorporate additional information. The revision makes use of a set of rules, which perform simple revisions, complex revisions and side transformations. The system that implements these principles is named STREAK, and it focuses on generating a single leading sentence of newswire reports.

Revision aims at adding additional information into an intermediate representation to generate more informative and concise descriptions. It combines representations and therefore can be taken as aggregation. New information can fit in at different depths including the sentence and phrase levels. At the sentence level, revision is the same as parataxis, and at the nominal level, both parataxis and embedding are possible.

The revision rules operate on a three-layered intermediate representation, which consists of a Deep Semantic Specification (DSS), a Surface Semantic Specification (SSS) and a Deep Grammatical Specification (DGS). A drafter produces an initial Layered Specification LS = <DSS,SSS,DGS> from the macro-structured DSS and then a reviser repeatedly applies revision rules to the LS to incrementally add information into the specification.

This idea of an intermediate representation is similar to that behind Meteer's Text Structure. Robin argues that Meteer's Text Structure is too abstract for the purpose of sports report generation because such a task involves achieving goals like conciseness and readability, which directly depend on surface form and cannot be achieved at a higher layer. His layered representation specifies more detailed syntactic and lexical constraints on the sentences that can be generated.

The LHS of a revision rule contains an LS pattern and an ADSS (Additional DSS) pattern, and the RHS is a list of revision operations for structure manipulation. All the ADSSs to be included are stored in a priority stack in an ordered manner. During each revision, the ADSS on top of the stack and the current LS form a pair to match the LHS of a revision rule, and then the RHS of the rule is executed in order. This is realised in such a way that no multiple rule matching is possible.

The revision rules cover some types of aggregation, e.g., coordination, and also other

phenomena such as pronominalisation. In order to avoid multiple rule matching, Robin devises a large set of specific rules, some of which use domain information, e.g., rules for adjoining frequency and game result data at the sentence level. Therefore, it is difficult to export such rules to a dramatically different domain like museum descriptions. In addition, it is not clear how well the revision approach is for generating texts with multiple sentences, where the readability of the text as a whole rather than just that of a single sentence needs to be considered during revision.

Unlike previous work on aggregation, Robin addresses the necessity of facilitating syntactic constraints on revision. His layered representation gives a way of constraining the syntactic complexity of the generated sentence.

Shaw and McKeown (1997) analyse a corpus of discharge summaries for patients in the medical domain to identify heuristics for hypotactic aggregation, which adds modifying constructions, e.g., adjectives and prepositional phrases, to predicative NPs. This is equivalent to embedding in our definition, but they do not consider referring expressions. The aggregation algorithm makes use of these heuristics and also needs to look ahead to the linguistic resources used by the surface sentence generator, including both lexicon and grammar, to determine whether a word or syntactic construction is available for combining propositions. This gives another way of providing lexical and syntactic constraints to aggregation.

Simple domain specific aggregation rules are also used in the following work. (Mellish and Evans, 1989) has a message optimisation stage, which makes use of domain specific rewriting rules to combine and simplify messages. The EPICURE system (Dale, 1990) involves simple discourse-level optimisation on the tree-structured discourse specification produced by the discourse planner to allow the clause generator to make use of conjoined verbs. In (Horacek, 1992), aggregation is integrated in the text structuring process by performing text structure (in RST style) modification.

(Hovy, 1993) argues that aggregation rules formulated in terms of discourse structure (e.g., an RST tree) can significantly reduce the complexity of the aggregation process. Such a rule looks like: "If two instances of the same RST relation emanate from a single Nucleus, then merge the two instances into one relation, and merge their Satellites into

the same leaf node."

As Hovy says, devising some rules for aggregation is not difficult, but the problem is what general rules exist. Researchers normally use their own intuitions in corpus analysis to derive rules. There is no discussion about how reliable these rules are. At the current stage, all three types of rules are used as a complement to one another in many rule-based NLG systems, where there is no clear boundary between the different types of rules.

2.4 Similarity-Based Aggregation

Aggregation can also be performed without the presence of explicit rules, with similar guidelines encoded in algorithms.

While much research performs aggregation on some sort of text structures using RST or other discourse relations, (McKeown et al., 1994) chooses a different approach for the PLANDoc system, which generates a 1-2 page summary of interactions between planning engineers and a telephone network planning tool. The Content Planner accepts a list of messages (in the form of semantic functional descriptions), which forms the whole text, and determines the ordering of these messages and the sentence boundaries. The ordering is based on a similarity measurement between messages and conjunction is used to express these with the maximum number of common attributes. To produce compound sentences, the planner first indicates the common and distinct attributes for each message and which common attributes should be gapped, then suppresses the gapped constituents. The aggregation algorithm of PLANDoc is described in detail in (Shaw, 1995).

(Shaw and McKeown, 1997; Shaw, 1998a) develop the above idea and implement aggregation as a part of the micro-planner of the MAGIC system, which generates information on patients in intensive care. In this system, the micro-planner is an extra layer between the content planner and the sentence generator. It takes as input a list of propositions in predicate-argument structures representing the content of a text, where each predicate or argument is elaborated by a number of feature-value pairs. Each step of aggregation enriches the argument representation, i.e., adds more feature-value pairs

to an argument. A complex representation can later be realised as a complex sentence.

(Shaw and McKeown, 1997) describes three types of aggregation: *semantic, hypotactic* and *paratactic* aggregation. Semantic aggregation changes the semantic structure of the central proposition. It mainly includes *nominalization* and *ontological subsumption.* Nominalization changes the predicate of a proposition so that a predicative NP carrying modifiers can be produced later on. For example, nominalization is used to produce "the patient is a male" instead of "the patient is male". Ontological subsumption is similar to Wilkinson's *conceptual* aggregation and it substitutes a set of concepts with a more general concept subsuming all those in the set in the ontology. For example, "cardiotonic therapy" is used instead of references to all components of the therapy. So this semantic aggregation covers phenomena not normally taken as aggregation by others and is more or less application specific. We have described the hypotactic aggregation in Section 2.3.3 and the paratactic aggregation has the same meaning as ours.

(Shaw, 1998b) focuses on *segregatory* coordination and presents a detailed algorithm for multi-proposition coordination, which has four steps: grouping and ordering semantic representations, detecting recurring elements, determining sentence boundaries and deleting recurring elements. In particular, the process of deleting recurring elements is driven by linguistic principles. Examples from the linguistic literature are used to show the capability or coverage of the algorithm.

Shaw combines a rule-based and a similarity-based approach in his sentence planner (named CASPER) and uses them for different types of aggregation. CASPER can handle more types of aggregation and also more complex aggregation than previous work.

In the PLANDoc and MAGIC domains for which Shaw's aggregation algorithms are designed, conciseness and being informative seem to be the major, if not the only, concerns of text structuring. For example, in PLANDoc, content planning involves only ordering and paraphrasing. In both systems, other structuring considerations are given such a small importance in the whole generation process that aggregation as a later process is not under many restrictions. It can reorder or change propositions to

facilitate the maximum amount of aggregation.

This approach will have difficulty in a domain where more important rhetorical goals need to be achieved through the selection and positioning of each proposition and being concise is only a secondary goal. In such a domain, it is potentially dangerous to redo earlier decisions because such operations might result in serious destruction to the satisfaction of some goals. Therefore, it is not always possible for aggregation to search for maximum similarity under no restriction. There is often a need to balance different text structuring considerations. This will become clearer through the discussion in Chapter 6.

Bateman et al. (1998) aim at designing an integrated architecture for both graphical and textual presentation to provide overviews of given datasets. This is realised in the KOMET-PAVE multimedia page generation system. To find out the regularities in the data to be expressed, they construct a dependency lattice which captures the redundancies in the data relations. This dependency lattice can then be used for both selecting graphic elements and generating text. For the textual aspect, the dependency lattice extracts partial commonalities in the data, which resembles the problem of aggregation in NLG. In fact, it represents all possible aggregations and each node of the lattice gives a point of aggregation. However, the lattice itself does not determine how to choose among aggregation possibilities. The decision is driven by the communicative intention. Therefore, no specific aggregation rules are needed and aggregation is considered "as a general property of all levels of linguistic representation constructed during the generation process". (Bateman et al., 1998) also shows how this strategy achieves the same effect as those using specific aggregation rules, e.g., (Dalianis and Hovy, 1996).

Although Bateman et al. (1998) mention incorporating aggregation into the larger picture of text structuring, the discussion is at a fairly abstract level. It is not clear how the decision for choosing an aggregation possibility is made. Besides, the dependency lattice is more useful if there are fair degrees of similarity among features of domain objects and there is a need to address these features together. This is not the case for our jewelry description domain, where objects are normally described individually.

(Robin and Favero, 2000) introduces the HYSSOP system, which summarises on-line analytical processing and data mining discoveries into hypertext reports. One major feature of HYSSOP is the way it performs aggregation. It is only concerned with parataxis, which works on deep semantic representations. The aggregation algorithm first tries to discover similarities in data sets using a generic sorting algorithm so that data cells with identical features are placed next to each other. It then generates a discourse tree representing a hypertext plan. Aggregation is encapsulated in the sentence planner and works under the restrictions of high-level textual organisation because the factors to be sorted are decided by the discourse planner. The output makes use of such layout features as bullet-points to enhance readability. This again presents a new application for text generation, where aggregation plays an important role.

The above discussion shows that a similarity-based approach is more appropriate for parataxis than for other types of aggregation. It suits a domain where being concise and informative are the main goals.

2.5 What Is Missing - the Interactions

The work we have introduced addresses in different depth the problems identified by (Reape and Mellish, 1999). To achieve a concise text, aggregation usually happens as a part of sentence planning, e.g., (Shaw, 1998a; Robin, 1994b), on different representations from text structuring to combine representations and sometimes fulfill other domain specific tasks such as nominalisation. However, the three additional problems we have identified are somewhat ignored in the literature. In particular, discussion about the relationship between aggregation and other NLG tasks is missing except that (Wilkinson, 1995) uses the relationship as a counter-argument for aggregation being a separate research topic.

Wilkinson argues that "aggregation may occur in such a wide variety of ways during generation that it seems impossible to isolate it as a unitary process." Firstly how to aggregate is just a special case of the general problem of sentence structuring; then the possibility of lexical aggregation makes it necessary to take into account possible

lexicalisation; and finally "there seems to be no aspect of language generation which can be excluded from a thorough consideration of aggregation." So he proposes to treat aggregation as a result of careful sentence planning. Indeed, as we have just mentioned, aggregation is often treated as a part of sentence planning in existing NLG systems.

Wilkinson's proposal implies that the phenomena handled by existing aggregation techniques can be covered by sentence planning, but we doubt that they can be handled only at the sentence level. There is also doubt whether the current sentence generation systems can deal with such complex phenomena as aggregation (Reiter, 1995). Therefore, we intend to take aggregation as a self-contained task, the task of achieving conciseness through combining representations, rather than a self-contained module or process. This task has to be considered in the context of satisfying other more important text structuring goals, that is, the goal of achieving a coherent text as a whole should be balanced against the consideration for conciseness. We believe that studying aggregation will help us to find a way to balance different planning considerations, which is important to the generation of a coherent text.

To satisfy multiple generation goals, the interactions between different phenomena must be modelled. In addition, a task can usually be achieved by a collection of processes, which have to be coordinated among themselves as well as with other processes. This explains the necessity of studying the interactions between aggregation and other generation processes. However, discussion in this respect is far from adequate. This can be shown by reviewing if and how the following relationships are addressed in the literature.

Aggregation and Document Structuring Since aggregation normally operates on adjacent propositions, the ordering of propositions, which is a problem of document structuring, can facilitate or block aggregation possibilities. Therefore, these two tasks are closely related.

Some research intends to change the decisions of text structuring to enable more aggregation. Dalianis calls this the *clause ordering problem* of aggregation (Dalianis, 1995; Dalianis, 1997b). His aggregation process reorders utterances to facilitate better ag-

gregation. Reordering of propositions is also a necessary part of (Shaw, 1998b) where aggregation is realised by a four-stage algorithm and the first stage is grouping and ordering semantic representations.

In STREAK (Robin, 1993; Robin, 1994b), content selection, organisation and realisation are somehow mutually constrained through the *layered specification*. However this mainly happens in *micro-level content organisation*, where drafting and revision of each base sentence are carried out. There is a clear division between macro-level and micro-level organisation, but no discussion about the effect of micro-level revision on the macro-level structure of the generated text can be found. In fact, STREAK does not have the option of producing several sentences. However, STREAK is only one example of the revision-based architecture. How well this architecture works in general remains to be seen.

The above approaches suit a domain with abundant regular patterns of expressing information. It is not clear how they work in a domain demonstrating less regularity but more flexible ways of expressing things, or when being informative is only one concern of the domain and it has to be coordinated with other more important communication goals like expressing interesting relations and generalisations between domain concepts.

In an experiment of (Dalianis, 1997a), Dalianis found that the shortest text was not necessarily the most readable text and coherence measures based on RST should also be considered. The common property of the work of Dalianis, Shaw and Robin is that no rhetorical planning is taken into account and therefore no discussion about how aggregation interacts with other coherence features is given, that is, the more difficult problem of achieving conciseness and rhetorical coherence at the same time is not addressed.

Subtypes of Aggregation It has been argued that aggregation is a multi-process task. In implemented systems, e.g., PLANDoc, the aggregation submodules are generally positioned one after another in the micro-planner. For example, in (McKeown et al., 1997; Shaw and McKeown, 1997), aggregation is accomplished through performing *semantic, hypotactic* and *paratactic* aggregation sequentially.

Rule ordering is itself an interaction problem between subtypes of aggregation. The most detailed discussion about this is given in (Dalianis and Hovy, 1996; Dalianis, 1997b), as introduced in Section 2.3.3. However, these orderings are mostly based on intuition. We do not have any idea of the optimal order of applying aggregation rules.

Aggregation and Referring Expression Generation Both embedding and parataxis aim at producing complex NPs, but there is no discussion in the literature about how they interact with the referring expression generation process which determines NP forms and restrictive modifiers. Some research simply performs referring expression generation after aggregation, e.g., (Shaw and McKeown, 1997; Shaw, 1998a). In (Shaw and McKeown, 2000), the quantification algorithm works on a set of predicate-argument structures where properties to be used for identifying the entities in them have already been decided by the referring expression module. The interaction between the quantification and referring processes is briefly mentioned but no further discussion is given.

Aggregation and Lexicalisation The so called Lexical Aggregation (Dalianis, 1996; Dalianis, 1997a) is a straightforward example of the interaction between aggregation and lexicalisation. This interaction is addressed in (Horacek, 1992; Wilkinson, 1995). We will not pursue this problem further in this thesis.

The above discussion shows that the study about the interactions between aggregation and other generation tasks is far from adequate. This is what we want to focus on in this thesis. Through studying the interactions, we wish to gain a better understanding of the problem of aggregation and possibly generation as a whole.

2.6 Summary

This chapter reviews the research on aggregation as to how it addresses the problems discussed in Section 1.1.3. We mainly look at how each work addresses such issues as which types of aggregation are modelled, what is the location of aggregation in generation, how aggregation rules are devised and what is the order of applying rules.

This helps us to identify the aspects of aggregation that need more work, which concern mainly two problems:

Interactions between aggregation and other processes : We have mentioned that in most systems, aggregation is carried out between text structuring and sentence realisation, i.e., in the micro-planner, as operations on an intermediate representation produced by the text planner. This representation is sometimes a list of propositions and sometimes an RST tree which captures the semantic connections between spans of a text and delimits the scope of the combination operation. In either case, neither the interactions between aggregation and other processes nor those between subtypes of aggregation can be modelled.

This thesis studies the interactions in detail, with a focus on the phenomena surrounding embedding. It presents a new way of abstracting and capturing the interactions, which makes it possible to perform aggregation as a part of text structuring.

Generality of aggregation rules : In implemented NLG systems, corpus analysis is often used to select general linguistic and psycholinguistic rules useful for the target domain as well as to devise complementary aggregation rules. Researchers often use their own intuitions in this process, so it is not clear if such intuitions can be shared by other people. We use more reliable means to devise embedding rules and principles in addition to using general linguistic observations, which include analysing a reliably annotated corpus and performing psycholinguistic experiment using multiple subjects. That is, we emphasise the sound empirical basis of the devised aggregation rules.

As to the representation aggregation works on, we use a similar method to (Robin, 1994b; Shaw and McKeown, 1997), i.e., requiring linguistic constraints in making aggregation decisions. We choose to use a revised version of Meteer's Text Structure (Section 7.2), which contains semantic and abstract syntactic information and can be used for document structuring as well as for aggregation.

From the next chapter, we will start to address the above problems from the perspective of embedding in NPs.

Chapter 3

Embedding in Referring Expressions

This chapter studies the role of embedding in generating complex referring expressions. We divide the components of a referring expression into a referring part and a non-referring part and discuss the complex interaction between the two parts. These motivate a set of syntactic and semantic constraints on the generation of the non-referring part. We use some corpus examples to illustrate the diversity and complexity of non-referring modifiers in museum descriptive texts.

3.1 Introduction

We have mentioned in the previous chapters that this thesis is mainly concerned with embedding, which makes decisions about certain types of modification in constructing complex NPs. In particular, we focus on the non-restrictive modifying components in referring expressions (RE) in the sense of (Kronfeld, 1990):

> "The term referring expressions are for those instances of noun phrase usage that are intended to indicate that a *particular* object is being talked about. Thus, whether or not a particular noun phrase is a referring expression depends on the way it is intended to be interpreted."

In the above description, the word *particular* excludes generic references, which refer to types of objects. Referring expressions can be NPs of various syntactic forms which refer to specific objects and they are very important and complex constructions in languages.

In English, referring expressions can be classified into definite and indefinite descriptions. Definite descriptions include proper nouns, pronouns and REs headed by possessive determiners or the determiner *the*. We are particularly interested in the last type of description because these pose the most difficult cases for embedding. In this thesis, we use definite descriptions to mainly refer to referring expressions headed by the determiner *the*. Indefinite descriptions are REs headed by the determiner *a*.

Research on RE generation focuses on deciding syntactic forms and choosing disambiguating modifiers (e.g., (Dale, 1992; Horacek, 1995), more discussion in Section 3.4). It seldom considers other types of modifiers. Work on aggregation is satisfied with devising a few rules to allow some degrees of embedding, rather than giving it an in-depth discussion.

In this chapter, we first divide the components of a referring expression into a referring part and a non-referring part, and give some examples of non-referring modifiers from our corpus of museum descriptions. We then discuss the relation between the two parts in some detail, which motivates the restrictions on the generation of the non-referring part, i.e., rules for embedding. We argue that generating non-referring modifiers is not an arbitrary or trivial decision. Because of the mutual restrictions between the two parts of an RE, the interaction between the referring expression generation task and the aggregation task, in particular embedding, is complicated.

3.2 An Analysis of Referring Expressions

To illustrate which part of a referring expression we study, we need a clear picture of the composition of a referring expression.

3.2.1 The Components of a Referring Expression

In addition to its primary function of denoting a discourse entity, a referring expression can serve other communicative goals such as providing new information about the entity and expressing the speaker's emotional attitude towards the entity (Appelt, 1985a). In Example (3.1), the underlined part refers to an object in a museum, and the part in boldface provides additional information about the object.

(3.1) *This example* ***from the time of the Qianlong Emperor 1736-95****, is made of lacquered wood with decoration in gold and red.*

We divide the components of a referring expression into two parts because they serve different functions/communicative goals and the rules for their content determination and realisation are different. The two parts are:

- *a referring part*: intends to refer to an object, but not necessarily to identify, that is, the expression denotes an individual object of a certain class, but it might not be necessary to know the exact object. The underlined part in Example (3.1) is a referring part.
- *a non-referring part*: intends to provide additional information about the referent denoted by the referring part, e.g., the part in boldface in Example (3.1).

This division is a functional one rather than a syntactic one. Except for the head and the determiner, which are always members of the referring part, other syntactic slots can belong to different parts in different circumstances. A referring part mainly serves the referring function, but it may also inform the reader about some properties of the referent. A non-referring part only serves the informing function and it is optional in a referring expression. In some sense, this division is similar to the restrictive/non-restrictive distinction, where the referring part equals the restrictive component and the non-referring part equals the non-restrictive component.

However, this division does not seem to work very well for indefinite descriptions. One reason is that the role of an indefinite is rather controversial and some people do not see it as a referring expression. Other people like Kronfeld (1990) argue that indefinite

descriptions can serve as referring expressions, and in most cases, there is a concrete entity corresponding to such a description, but the use of an indefinite description normally signals that the identity of the referent is not important. This means that indefinites can be used to refer but not to identify. In this thesis, when we mention indefinite phrases, we mean the second interpretation of indefinite descriptions, unless they are otherwise noted.

Another reason is that an indefinite is normally used for the first mention of an object and therefore all properties in the expression are new information. The distinction between the information for referring and informing in this case is less obvious than that in a subsequent reference.

In Section 3.2.3, we give some examples from our corpus of museum descriptions (to be introduced in Section 4.2) to illustrate the complex REs in human written texts. When choosing these examples, we had to use an intuitive criterion to look for optional information in the referring expressions, that is, we tried to find the minimal description that is necessary for understanding a sentence by removing a piece of information from the expression and judging by intuition if the meaning of the sentence is changed dramatically. If not, the information is non-referring. For example, in (3.2a) we think *unruly* is a referring property because its presence is essential for the meaning of the sentence. In (3.2b), the prepositional phrase and relative clause in boldface are obviously informing, whereas *black* can be both.

(3.2) a. *It describes events connected with the Chou King's campaign to discipline an* ***unruly*** *vassal, P'u-tzu, the ruler of a state south of the Han River in Hupei.*

b. *The text is circumscribed by a* ***black*** *border* ***of 22.8 × 16.2 cm, which consists of double lines on the right and left sides of the page.***

One difference between text analysis and generation is the resources available to them. In analysis, we directly face the words and phrases which compose human languages, whereas in generation what we have is the world knowledge organised under certain principles. In order to demonstrate the types of property expressed as a non-referring modifier, we need a conceptual ontology to abstract away from the concrete words/phrases in the corpus. In the next section, we describes such an ontology: the

Generalized Upper-Model.

3.2.2 The Upper-Model Classifications of Predicates and Modifiers

The Generalized Upper-Model (GUM) (Bateman et al., 1995) is a hierarchical organisation of the concepts (i.e., things, processes, properties, etc.) that may be expressed in languages. It has been used in a number of generation systems, e.g., Penman (Mann, 1983) and KPML (Bateman, 1995).

The GUM has a taxonomy for predicates. In theory, each NP modifier can be mapped to a concept in the predicate ontology, although ambiguity cannot always be avoided. The fragment that is most relevant to embedding is given in Figure 3.1.

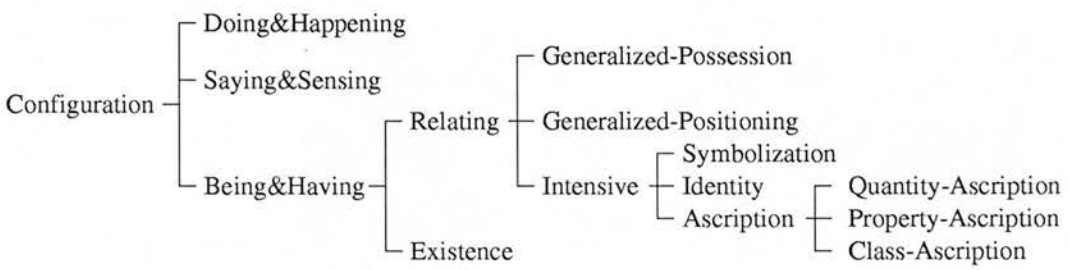


Figure 3.1: A fragment of the Generalized Upper-Model

A proposition whose predicate is subsumed by *Property-Ascription* can usually be realised more briefly as an adjective of an NP. So we consider the semantic feature of an adjective as *Property-Ascription*. However, it is not very useful to classify all propositions describing object properties into one class. We need a more refined classification for adjective modifiers.

We used the quality ontology of GUM to extend *Property-Ascription*. In the Upper-Model, qualities of objects are called *Material-world-qualities*, and there are five types of qualities:

Status-quality: a quality ascribed to an object and independent of the observer, e.g., *dead/alive*.

Class-quality: a quality of being made of a particular material or a quality of place, institution, social group, or other social category of origin, e.g.,

wooden, English.

Sense-and-measure-quality: a quality that is sensed or measured by conscious beings, e.g., *young/old, expensive/cheap.*

Evaluative-quality: a quality that is determined by some value system of some conscious beings, e.g., *honest, beautiful.*

Behavioral-quality: a quality that characterises the behaviour of a conscious being, e.g., *clever, enthusiastic.*

Each quality type gives *Property-Ascription* a new subtype, e.g., *Evaluative-quality-ascription.* In this way, we get a more refined classification of the adjective modifiers in our corpus.

We used the above classification and terminology to classify modifiers in our corpus. The examples in the next section demonstrate different types of property expressed as non-referring modifiers.

3.2.3 Examples of Non-referring Modifiers

We are concerned with how many types of referring expression there are and what additional information is usually expressed in them in museum descriptions. We found that the non-referring parts were of surprising diversity, more than had been considered in existing NLG systems (except for STREAK). In this section, we will illustrate them through examples from the selected domain texts (details in Section 4.2). The examples are organised according to the syntactic forms of the modifiers and the NPs.

Adjectives: Evaluative-qualities

Demonstrative - this/these:

(3.3) *Records of similar ceremonies are almost non-existent in historical texts, and it has been suggested that it is because some emperors had used these* ***extremely costly*** *occasions not for the benefits of their subjects, but to pursue selfish ends.*

Definite - the:

(3.4) *The* ***solid yet elegant*** *characters on these tablets were based on the actual writing style of Emperor Xuanzong.*

Definite - possessive:

(3.5) *He was executed in 1661 for his* ***treasonable*** *involvement with those who had rebelled against Charles I.*

Indefinite:

(3.6) *In this room, wood panelling and a reconstructed plaster ceiling from the house of a* ***prosperous*** *burgess in Kirkcaldy give an impression of how a well-off family lived.*

Adjectives: Sense-and-measure-qualities and Status-qualities

We do not distinguish the two because the distinction between them is very vague.

Demonstrative - this:

(3.7) *This* ***broad, deep*** *bowl with flaring rim and convex bottom is supported by three slightly curved triangular legs.*

Definite - the:

(3.8) *The* ***cylindrical*** *body of this tsun is divided horizontally into a slightly flared foot, a swelling midsection, and a widely flaring mouth.*

Indefinite:

(3.9) *The cylindrical body of this tsun is divided horizontally into a* ***slightly flared*** *foot, a* ***swelling*** *midsection, and a* ***widely flaring*** *mouth.*

Prepositional phrases

Demonstrative - this:

(3.10) *This broad, deep bowl* ***with flaring rim and convex bottom*** *is supported by three slightly curved triangular legs.*

Definite - the:

(3.11) *The throne cover,* ***from the reign of Jiaquing, 1796-1820,*** *is woven in yellow silk which is the imperial colour of the Qing Dynasty, 1644-1911.*

Definite - possessive:

(3.12) *After the birth of her daughter Mary and her husband's death* ***in 1542,*** *Mary of Guise wanted to be close to the centre of power.*

Proper name:

(3.13) *The banner takes its name from Fetternear,* ***near Aberdeen,*** *where it was found in the 19th century.*

Indefinite:

(3.14) *On one side is a strap handle, and the lip is surmounted by two upright columns* ***with mushroom-like caps.***

Apposition

Definite - the:

(3.15) *The Yi,* ***an ewer for washing the hands,*** *was associated with the p'an used in ritual ablutions, according to Tso Chuan.*

Proper name:

(3.16) *Mary of Guise,* ***second wife of James V and mother of Mary, Queen of Scots,*** *may have lived in the house between 1542 and 1554.*

Indefinite:

(3.17) *It describes events connected with the Chou King's campaign to discipline an unruly vassal,* ***P'u-tzu, the ruler of a state south of the Han River in Hupei.***

Non-restrictive clauses

Demonstrative - this:

(3.18) *Such virtuosity, subtlety, and attention to the details of expression have been unmatched by any of Kuo Hsi's followers, and this monument,* ***dated 1072 A.D.,*** *stands as the painter's supreme masterpiece.*

Definite - the:

(3.19) *His hands make the gesture of preaching,* ***known in Sanskrit as the dharmacakra-mudra.***

Definite - possessive:

(3.20) *It is remarkable for its craftsmanship and colour* ***which after 500 years has faded only slightly.***

Proper name:

(3.21) *The solid yet elegant characters on these tablets were based on the actual writing style of Emperor Xuanzong,* ***who was also a well-known calligrapher of the Tang dynasty.***

Indefinite:

(3.22) *The text is circumscribed by a black border of 22.8 x 16.2 cm,* ***which consists of double lines on the right and left sides of the page.***

Nouns/Compound nouns before the head

These are usually referring, but can inform sometimes.

(3.23) a. *The panelling's* ***red pine*** *timber was probably imported from the Baltic, a key trading destination for merchants on the Fife coast.*

b. *The scene strongly resembles an illustration entitled K'uei-chien t'u, which was originally designed by the* ***late Ming period*** *painter, Ch'en Hung-Shou, for inclusion in a wood block print edition of Hsi-hsiang chi "Romance of the Western Chamber".*

3.3 The Relation Between the Components of a Referring Expression

Although the referring part and the non-referring part serve different communicative goals (the former intends to refer to an object, whereas the latter provides additional information about it), they are closely related to each other.

On the one hand, the referring part puts both syntactic and semantic constraints on the presenting of the non-referring part. The syntactic constraint concerns mainly the available syntactic slots around the head. For example, if the referring part has a relative clause for identifying the referent, it would be better not to add other post-modifiers as otherwise the sentence might be too complex to interpret or have attachment ambiguity; or if the referring part is a pronoun, it would not be possible to add any new information.

The semantic constraint is that if the referring part can uniquely identify the referent, the reader should not be confused over which object the referring expression is about because of the addition of the non-referring part. For example, in the description of a current focal object which is a necklace, we might say (3.24a) below. Suppose we also want to inform the reader that the necklace has floral motifs. We should use (3.24b) rather than (3.24c) because (3.24c) may make the reader think that the sentence is about a necklace which is not the focal object.

(3.24) a. *The necklace is made of sapphire, enamel and gold.*

b. *The necklace,* ***which has floral motifs****, is made of sapphire, enamel and gold.*

c. *The necklace* ***with floral motifs*** *is made of sapphire, enamel and gold.*

d. *This necklace **with floral motifs** is made of sapphire, enamel and gold.*

If an object is in the immediate situation of an utterance, i.e., visibly salient referent as assumed in (Hawkins, 1978), a demonstrative phrase can be used instead of a definite one, and no confusion will be caused by adding any new information, e.g., (3.24d). That is, the referring part restricts the presentation of the non-referring part semantically through the way the referent is realised.

On the other hand, the possibility of adding a non-referring part can make some realisations of a referent preferred over others. A referent can usually be realised in a number of ways and linguistic research suggests certain preferences among possible realisations. For example, Gundel et al. (1993) present a *Givenness Hierarchy* consisting of six cognitive statuses and their corresponding NP forms, which looks like:

> *in focus (pronoun)* $\prec$ *activated (demonstrative description)* $\prec$ *familiar (that N)* $\prec$ *uniquely identifiable (definite description)* $\prec$ *referential (indefinite this N)* $\prec$ *type identifiable (indefinite)*

Each status entails all lower statuses, that is, the speaker uses a particular NP form to signal that she assumes the associated cognitive status is met and thus all lower statuses to the right. For example, by using a *definite description*, the referent must be *uniquely identifiable* and therefore *referential* and *type identifiable.*

Gundel et al. (1993) also predict that a linguistic form can appropriately encode its corresponding cognitive status as well as all higher statuses. For example, a *definite description* can denote a referent *uniquely identifiable* or *in focus*, but not a referent that is only *type identifiable.* And they expect the forms to vary for one status in actual discourse. The predictions are validated by their discourse analysis based on five languages, although there is strong preference for some forms by some statuses. This work indicates that a cognitive status can be realised in multiple linguistic forms and a form can be used for multiple statuses.

A more specific proposal is about the realisation of the backward-looking center (Cb) in Centering Theory (Grosz et al., 1995), where *centers* of an utterance refer to "those entities serving to link that utterance to other utterances in the discourse segment

that contains it". A *Cb* links an utterance to the preceding discourse and is usually the most salient element in the previous utterance. If the *Cb* of the current utterance is the same as that of the previous one, pronominalisation of the *Cb* is often preferred as this signals the continuation of the same topic. However, except for the claim that a *Cb* should be pronominalised when a non-*Cb* in the same utterance is pronominalised, Centering Theory does not preclude using any other syntactic form such as a definite description for the *Cb* in other situations.

RE generation modules tend to choose a pronoun for a *Cb* in each center continuation. This would undoubtedly produce rather boring texts in the domain of descriptive texts because there are few variations in the subject NPs. It is still an open question why in human written texts, a *Cb* is not realised as a pronoun but rather a definite phrase, including a proper name, in many cases.

Grosz et al. (1983) observe that full noun phrases containing some new and unshared information can be used to refer to the current centered entity, but they argue that in this case additional inferences are needed by the readers to determine that the center has not shifted and that the properties expressed hold for the centered entity. Henschel et al. (2000) specifically mention that using identical repeated pronouns at the clause onset is rare in expository and descriptive texts (only 2.6% of all discourse pronouns in their corpus), and human writers often use various aggregation techniques to introduce variation into NPs. Henschel et al. think that the goal of blocking pronoun repetition triggers aggregation, which results in the apparent frequency of definite descriptions (including proper names) in their corpus.

So there is an obvious need to balance the considerations for local coherence (more details in Section 6.1.1) and stylistic preferences (DiMarco and Hirst, 1993) like avoiding repetitions in the subjects.

To take into account all constraints and preferences to some extent, more than one possible realisation of a referent, with no significant difference in the degree of coherence, may have to be considered. As a result, the one that is more suitable for adding new information can be chosen. In example (3.25), the continuity of the center *Jessie King* can be realised by either a short name, a definite phrase or a pronoun. The definite

expression is chosen in (3.25a) because of the added new information, as compared to (3.25b).

(3.25) a. *Jessie King designed this necklace. The **famous Scottish** designer worked for Liberty & Co.*

b. *Jessie King designed this necklace. King worked for Liberty & Co. She is Scottish. She is famous.*

In NLG systems, the referring expression generation process is usually implemented as a part of or just before surface realisation. The algorithm should ideally obey Grice's Maxims of Quantity, Quality, Relation and Manner for content selection and surface realisation. However, we expect different strategies to apply for the non-referring part as it serves different communicative goals. So we need two processes to generate a complex referring expression:

1. a referring process, which generates the referring part, e.g., (Dale, 1992; Horacek, 1997);

2. an embedding process, which selects suitable properties for the non-referring part and realises them as components within the structure of a referring expression. As already mentioned, this is a subtype of aggregation.

Because of the mutual effect on one another between the two parts of a referring expression, the two processes interact with each other in a complex way. In the next two sections, we will discuss them separately and in Section 7.3.3 we will try to coordinate them.

3.4 Generating the Referring Part

The first formal model of referring was presented by Appelt within the framework of a general theory of speech acts and rationality (Appelt, 1982; Appelt, 1985b; Appelt, 1985a; Appelt, 1987), where referring is treated as a speech act to establish mutual belief between speakers and hearers concerning the speaker's intention to refer to a particular object. The model can explain how referring acts achieve multiple goals

including referring, providing information about a referent, and requesting an identifying action to be performed. Yet there is a big gap between the general model and the actual planning of the linguistic content of a referring expression.

Later research follows a more practical track and a great deal of work has been done on generating various types of referring expressions. One of the most comprehensive lines of work in this aspect is done by Robert Dale (Dale, 1986; Dale, 1987; Dale, 1990; Dale and Haddock, 1991a; Dale and Haddock, 1991b; Dale, 1992; Dale and Reiter, 1994). In his early work, Dale discusses in detail various types of object referring expressions and how they are generated in the EPICURE system. He gives the principles of *sensitivity, adequacy* and *efficiency*, which must be obeyed by the referring process to generate an expression taking into account the hearer's knowledge, being sufficient to identify the intended referent and providing no more information than is necessary for the identification of the intended referent. In EPICURE, the referring function is realised in the clause generator and has two steps:

1. Determining the *recoverable semantic structure* of the NP describing a domain entity, using KB→RS (knowledge base to recoverable semantic) mapping rules;
2. Mapping the semantic structure to the *abstract syntactic structure*, using RS→AS (recoverable semantic to abstract syntactic) mapping rules.

The two steps correspond to NP content determination and realisation respectively.

(Dale and Reiter, 1994) examines the computational complexities of strictly obeying Grice's Maxims in RE generation, and proposes a faster and simpler incremental algorithm that resembles the behaviour of human speakers. The algorithm does not attempt to look for the "optimal" attribute set for the referring task, but simply iterates through a list of available attributes in a fixed order to include those with some discriminating ability. (Horacek, 1995; Horacek, 1997) revise this algorithm in a number of ways. Currently, various versions of Dale's incremental algorithm are being used in NLG systems, e.g., (O'Donnell et al., 1998).

3.5 Restrictions on the Non-referring Part

Little research addresses the generation issues with respect to the other functions of a referring expression. An exception is the work of O'Donnell et al. (1998), where NPs are generated to serve both referring and informing functions. However, their approach emphases the effect of adding informing properties on the choice of NP forms, but does not give centering enough prominence. Although Scott and de Souza (1990) discuss the relation between embedding and rhetorical relations and give several heuristics for combining sentences using embedding (see Section 2.3), the connection between embedding and referring expression generation is not mentioned.

Since an RE is primarily for referring to an object, other functions can only be fulfilled when they do not interfere with the primary function. To make sure that the non-referring part fits into an RE properly, we summarise three rules that the non-referring part must obey.

Rule 3.1 *The non-referring part should not confuse the reader about the referent indicated by the referring part.*

That is, if the referring part can uniquely identify the referent, the reader should not be confused about which object the referring expression is about because of the addition of the non-referring part. An example was given in (3.24), where (3.24c) is inappropriate because the added prepositional phrase in the subject might confuse the reader about the intended referent.

This rule requires that only properties or syntactic realisations obviously for informing rather than referring are used for the non-referring part. This can be achieved by both semantic and syntactic means, more specifically,

- NP type: how a referent is realised determines which extra information can be added into an expression. For non-pronominal phrases, determiners play a crucial role in distinguishing between referential and other properties. For example, the prepositional phrase *with floral motifs* is referential in *the necklace with floral motifs*, but possibly informative in *this necklace with floral motifs.*

- Realisation of a non-referring part: modifiers separated from their head by a comma are mostly informing.
- Property type: human generated referring expressions often demonstrate preferences toward using certain types of properties for object identification and some other properties for additional elaboration.

These will be further discussed and illustrated through corpus statistics in Chapter 4.

Rule 3.2 *The non-referring part should not reduce the readability of the text.*

By readability we mean the fluency of a text. The complexity of a referring expression affects the readability of a text because referring expressions are a major component of the text. This rule requires that the generated NP and the text as a whole should not be too complex to read. There are several restrictions concerning readability. Here we only discuss those affecting clause level comprehensibility and leave the others to Chapter 6.

It is obvious that complex referring expressions are more difficult to read. To avoid difficulties in the comprehension of a referring expression, (Scott and de Souza, 1990) gives a heuristic which allows only one level of clause embedding. Coleman (1962) argues that when people read a sentence, they try to rebuild the relation between the subject and the verb. If the distance between them is too big, the readability of the sentence is reduced. According to this observation, embedded clauses in subjects are a major obstacle to comprehensibility.

Complexity is closely related to the user model since different groups of people, children/adults or non-native/native speakers, would not have the same ability to understand complex sentences. A generation system should adjust the amount of information packed into a referring expression, especially that inside the subject, according to different user configurations.

Rule 3.3 *The non-referring part should not change the properties of the referent.*

Speaking at the surface level, one basic requirement on aggregation is to make sure

that, given a sequence of individual sentences, there is no dramatic change in the meaning expressed by the unaggregated and aggregated sentences. A special case for embedding is relevant to the different uses of a lexical item. For example, (3.26b) gives a different description about the boy than (3.26a).

(3.26) a. *The boy works in a restaurant. He is poor.*

b. *The* ***poor*** *boy works in a restaurant.*

The word *poor* can only be in a position before a noun when it means unlucky, whereas its positioning is more flexible when it means having very little money. If this piece of information together with usage preferences are present in the lexicon, embedding could take into account the different usage and avoid this side-effect. This requires the adoption of a similar method to that of (Shaw and McKeown, 1997), that is, looking ahead to the linguistic resources used by the surface sentence generator, including both lexicon and grammar, to determine whether a word or syntactic construction is available for combining propositions.

Rules 3.1 and 3.3 are related to the more general issue of avoiding ambiguity in aggregation. Aggregation is not just about joining things together, it has to prevent the undesirable side-effect this might have. As Horacek (1992) mentions, the goals of achieving conciseness and presenting information accurately must be balanced against each other. Work in this respect mainly concerns parataxis. For example, (Dalianis, 1996) uses cue words such as *together* and *each* to avoid ambiguities that might be introduced by parataxis. This is achieved by associating a cue primitive selection process with each aggregation rule. Also with the goal of producing more concise and less ambiguous texts, (Shaw and McKeown, 2000) gives an algorithm for generating quantified referring expressions to refer to sets of distinct domain entities. The algorithm first identifies the set of entities that can be quantified and then makes appropriate generalisation and selects a suitable quantifier according to a given set of axioms. However, there is no previous discussion about which kinds of ambiguity embedding can cause and how to avoid them.

Currently, there is no theoretical framework or effective algorithm to guarantee that the meaning is not changed or no unwanted meaning is introduced by aggregation. Luckily,

this seems to be less of a problem in our domain, which mainly provides descriptive information about domain objects. We will not pursue problems related to Rule 3.3 further in this thesis.

To generate the non-referring part properly, these three rules must be taken seriously. How to obey the rules is both a theoretical and an implementation issue. The essence of this thesis is to investigate the phenomena associated with conforming to these rules and try to find out how these phenomena can be captured by a generation system in a principled way. In Section 7.3.3, we describe an algorithm for generating referring expressions, intending to capture the interaction between the referring and non-referring parts. We also introduce Meteer's Text Structure in Section 7.2 to provide abstract syntactic constraints on text planning and referring expression construction. This gives an effective representation for controlling the complexity mentioned in Rule 3.2. Theoretical issues relevant to the other parts of Rule 3.2 will be discussed in Chapters 5 and 6. In Section 4.4, we discuss one of the most difficult cases with regard to Rule 3.1.

3.6 Summary

In this chapter, we first divide the components of a referring expression into a referring part and a non-referring part and discuss the complex interaction between the two parts. We then focus on the general rules and factors that affect the content determination and construction of the non-referring part, which are the tasks of embedding. We use examples from our corpus to illustrate the diversity of non-referring modifier usage.

Through the discussion, we wish to show that embedding in referring expressions is a complex decision not just sensitive to the grammatical issues in language, but also to the discourse context. Further support to our heuristics will have to come from corpus analysis and the evaluation of the final implementation.

Chapter 4

Corpus Analysis

This chapter describes two corpus analyses we performed to discover the regularities in the usage of modifiers in noun phrases. Our corpus is mainly composed of museum descriptive texts. The first analysis reveals the general characteristics of NP modifiers in such texts. The second analysis uses more systematic and fine-grained approaches depending on reliable annotation of the corpus with features of NPs and modifiers that might affect the decisions of modifier content and realisation. The valuable results from this analysis include figures concerning the additional information that is usually expressed through NPs and a decision tree for modifier type determination. These provide a reliable basis for our embedding rules and modifier generation algorithms described in the other chapters of this thesis. We also discuss the factors affecting the quality of embedding in definite descriptions headed by the determiner "the".

4.1 Motivation

In NLG, corpus analysis is often used to identify the linguistic resources that convey certain information, to motivate specific generation architectures or algorithms, or to collect data for evaluation. In other words, through corpus analysis, researchers try to guess what principles human authors might have in mind when they handle domain specific writing tasks and what generation strategies could be used to simulate their behaviour to achieve similarly coherent texts. For example, Robin (1994b) analyses a

corpus of reports for basketball matches, which leads to a set of revision rules and a revision-based generation architecture. These enable his system STREAK to generate reports bearing similar structures, such as containing a large amount of historical information, to those in the corpus. Dalianis (1997b) uses corpus analysis to show the frequency of aggregation and the use of cue words for disambiguating aggregated sentences in human-authored texts. However, previous corpus analysis mainly focuses on single sentences, e.g., (Robin, 1994b; Shaw and McKeown, 1997), so its results cannot satisfy our needs.

Through analysing museum descriptions, we wish to find out in a coherent text which information human authors typically convey in a referring expression or an NP in general and how they realise it. Since the descriptions we will study are coherent human texts, the NP modifiers in them are produced under both coherence and conciseness considerations.

However, the set of referring expressions (if such a set can be identified) is still too big for us to handle. We do not intend to generate all types of referring expressions found in our corpus or generate exactly what appears to be there. The corpus analysis is for suggesting general rules for embedding and a computational model which can simulate domain specific embedding phenomena.

This chapter describes two corpus analyses we performed. The first tries to characterise the referring expressions in a corpus of museum descriptive texts (details are given in Section 4.2). The results are used in the implementations described in Chapter 7.

The second analysis uses more systematic approaches such as corpus annotation to get more reliable evidence for suggesting embedding algorithms. This is a part of our effort in the GNOME project (Generating NOMinal Expressions (Poesio, 2000a)), which is a joint project between the University of Edinburgh and the University of Brighton. GNOME aims at the development and implementation of general algorithms for the generation of nominal expressions, which prove applicable to different NLG system environments. The central idea is to build statistical models for NP form determination by training on an annotated sub-corpus. Details about this analysis are given in Section 4.3.

Although this analysis produces some very interesting results, we did not use them in our implementations. The main reason is that the implemented systems ILEX-TS and GA-plan can only produce limited types of NP and sentence structure, therefore the difference in algorithms cannot be demonstrated through the system output.

The last section of this chapter studies the heuristics for embedding in definite descriptions with the definite article *the*, including possessive phrases, and partially validates the heuristics through corpus observations.

4.2 An Analysis of Museum Descriptions

Our corpus consists of museum descriptions in English, including:

- 15 from the Far East Collections (the IvyWu Gallery), the National Museum of Scotland (NMS), which contain historical items from different historical periods of China, Japan and Korea.
- 26 from the Scottish Culture and Scottish Medieval Collections, NMS, which contain items used in Scotland from the twelfth to the twentieth century.
- 5 from other parts of NMS.
- 24 from the National Palace Museum, Taipei, which houses precious treasures from thousands of years of Chinese history.
- 10 from the Louvre collection (des Musees Nationaux, 1993).

These 80 texts are selected from a much larger collection of texts according to the following principles:

- Each text describes concrete objects rather than abstract concepts, preferably a single object rather than a group of objects.
- Each text contains relatively complex referring expressions but simple verbal phrases.

4.2.1 General Characteristics of REs

Museum descriptions are normally short; most of them are less than 200 words. Descriptions from different museums vary considerably in the complexity of sentences and noun phrases. The longest sentence collected has 55 words, including four prepositional phrases, three non-restrictive clauses and two conjunction words.

From the 80 texts, we randomly selected 20 for detailed analysis. The non-referring part is the focus of our corpus analysis. We used our intuition to look for optional information in a referring expression (we tried to compensate for this use of intuition in the second corpus analysis – Section 4.3). We collected the referring expressions from these texts, among which 240 (64%) are definite descriptions and 135 (36%) indefinite ones. Table 4.1 illustrates the distribution of the non-referring modifiers in these phrases (the figures for referring modifiers are given for comparison).

Types	**Definite Descriptions**			**Indefinite Descriptions**		
	Refer	*Non-refer*	*Total*	*Refer*	*Non-refer*	*Total*
Adjective	15	32	47	8	63	71
Prepositional Phrase	69	12	81	38	16	54
Non-restrictive clause	0	23	23	0	33	33
Apposition	2	12	14	0	2	2

Table 4.1: The distribution of non-referring modifiers in the collected REs

We used the GUM classification and terminology introduced in Section 3.2.2 to classify these modifiers. We counted the occurrence of each Upper-Model concept mapped from a word or phrase in the corpus. If certain concepts appear significantly more frequently as non-referring modifiers than others, we can define embedding rules to capture this regularity. Here we summarise the characteristics of the referring expressions in these texts:

- Many referring expressions (30%) are simple, in the form of a proper name, a pronoun or a definite phrase without any modifier. Nouns before the head are usually for referring, with only rare exceptions. Complex referring expressions appear frequently in these texts, some of which are heavily packed with premodifiers and postmodifiers.

- In the definite descriptions collected, there are 47 adjectives, in which 32 are for providing additional information, according to our judgement. That is, 2/3 of the adjectives provide properties not for referring. Of these adjectives, all those describing *evaluative-qualities* and 2/3 of those describing *sense-and-measure-qualities* and *status-qualities* are for providing additional information. Adjectives describing *class-qualities* are usually for referring, but can also inform sometimes. In the indefinite descriptions collected, there are 71 adjectives, in which 63 give additional information.

  There seems to be no limit to the number of adjectives that can appear before the head. Based on an analysis of human written sentences, (Coates, 1977) states that sequences of two modifiers are far more common than sequences involving larger numbers. This is also the case in our corpus.

- Of the 81 prepositional phrases collected from the definite descriptions, only 12 provide additional information. Most of them denote the *generalized-possession* relation (by *with*) between two objects and a comma is often used to separate a prepositional phrase from its head. So prepositional phrases are usually for referring. Among the 54 prepositional phrases from the indefinite descriptions, only 16 are for informing.

- There are 23 non-restrictive clauses of various kinds and 12 appositions in the definite descriptions, and 33 non-restrictive clauses in the indefinite ones. There are far more non-restrictive clauses than restrictive ones. These non-restrictive clauses are usually in reduced forms (e.g., *-ing/-en*), and only a small proportion have explicit markers like *which* and *who*.

We also observed that the distribution of the additional information is uneven. Some referring expressions are packed with many modifiers, like those introducing a new object, whereas others are very simple.

4.2.2 Deriving Embedding Rules

The above observations suggest general rules to enable an NLG system to simulate embedding in human texts. For example, those facts describing evaluative qualities

are the best choice for embedding. This results in our first set of embedding rules (we do not consider the interaction between aggregation and other text generation modules at this point). The decisions are on two aspects:

- How much additional information can be expressed in a referring expression? This is determined by the realisation of the referring part.
- Which abstract surface form is suitable for a non-referring property? This is determined by the semantics of the fact to be embedded.

Without considering the complexity of a referring expression, the embedding rules related to the first aspect can be described as (examples from Section 3.2.3 are used to illustrate the rules):

1. Since the referring part of a demonstrative phrase can always uniquely identify the referent, any information can be added (Examples (3.3), (3.7), (3.10) and (3.18)). There are 13 (7%) modifiers in the collected demonstrative phrases and all of them are non-referring.
2. Since the referring part of an indefinite description does not intend to identify the referent, any information can be added (114 (59%) cases in our collection). The additional information is both restrictive and descriptive (Examples (3.6), (3.9), (3.14), (3.17) and (3.22)).
3. It is very rare to use premodifiers (0 case) for a proper name in human written sentences, but postmodifiers such as appositions and non-restrictive clauses can be used naturally (Examples (3.13), (3.16) and (3.21)). There are 24 (12%) such cases in our collection.
4. Since both the head and the modifiers in the referring part of a definite description are to uniquely identify the referent, the properties that can be embedded are restricted to those that are obviously informing. This includes three cases:
    - Evaluative premodifiers seldom refer, so they can be expressed in a definite description (Example (3.4)). 9 (5%) such modifiers were collected.

- Postmodifiers that are separated from the head by a punctuation mark like a comma, a dash or a bracket are obviously informing, so any information can be added this way (Examples (3.11), (3.15) and (3.19)). We collected 37 (19%) such modifiers.
- Other premodifiers can be added according to the context or the realisation of the referring part (Examples (3.8) and (3.23b)). This will be discussed in detail in Section 4.4.

Possessive phrases belong to this category in our analysis (Examples (3.5), (3.12), (3.20) and (3.23a)).

Since a modifier can be classified to more than one of the above cases, for example, a modifier can be realised in a demonstrative phrase and separated from the head by a comma, the total percentage is larger than 100%.

For the second aspect, suppose the fact to be embedded is *predicate(argument)*, where *argument* represents a discourse entity. We want to show how this semantic representation relates to its abstract surface realisation in a referring expression. We are particularly interested in those properties that can be realised as a component other than a non-restrictive clause. The rules are illustrated in Table 4.2. The first column gives the Upper-Model concept subsuming *predicate*, which decides the abstract surface realisation of the fact. It also implies that such information is more likely to be embedded. The numbers show how many (percentages of) non-referring modifiers in our collection are covered by each rule. In the table, *Property-Ascription* covers the more specific categories such as *evaluative-quality-ascription* and *status-quality-ascription.* The third column gives the surface forms of the embedded facts in some examples in Section 3.2.3 and also gives a new example for *Class-Ascription*, which displays the desirable embedding. The table does not intend to give a complete list of possible embedding types and more flexible phenomena often appear in human written texts.

The above rules cover 90% of the non-referring modifiers in our collection. The rest is mainly semantic concepts realised as prepositional phrases.

UM cat. of *predicate*	**Abstract surface form**	**Presentation**
Property-Ascription (95, 49%)	Evaluative premod	(3.4): The characters are solid. They are elegant.
	Sense-and-measure/ status/class premod	(3.8): The body is cylindrical. (3.9): The foot is slightly flared.
Class-Ascription (14, 7%)	Head/Apposition	Clad in his trademark black velvet suit, the soft-spoken ***clarinetist*** announced that his new album had just been released.
Generalized-Possession (10, 5%)	Prepositional postmod	(3.14): The columns have mushroom-like caps.
All others (56, 29%)	non-restrictive clause	(3.18): This monument dated 1072 A.D.

Table 4.2: First set of embedding rules

For facts whose predicates are of the *isa* type, additional information is expressed through adverbial phrases rather than being embedded into the referring expressions. For example, Sentence (4.1a) is normally used instead of (4.1b).

(4.1) a. *The banner is a unique survivor,* ***the only known church banner from medieval Scotland.***

b. *The banner,* ***which is the only known church banner from medieval Scotland,*** *is a unique survivor.*

This corpus analysis gives us some general impressions about the non-referring modifier usage in REs. The set of rules gives a way to conform to Rule 3.1 discussed in Section 3.5. However, there are a few problems with the analysis. Firstly, we found that the referring/non-referring distinction suited generation better than analysis and in quite a few cases, it was difficult to make the distinction, e.g., for indefinite descriptions. This was also the case for classifying a non-referring modifier using Upper-Model concepts, whose definitions were not always clear. It was often difficult to distinguish between concepts. Secondly, we depended on our own intuition to classify modifiers. This intuition might not be shared by other people.

To obtain more solid evidence for embedding algorithms, a more systematic and fine-grained analysis of corpus texts is needed. This leads to the second corpus analysis.

4.3 An Annotation-Based Corpus Analysis

The second corpus analysis is a part of our work on NP modifiers in the GNOME project (Poesio, 2000a). The GNOME corpus consists of two parts: museum descriptions and patient information leaflets, which give instructions on how to use certain drugs. Our work in GNOME studies all types of NPs in the corpus rather than just referring expressions.

4.3.1 Refinement of Features

A major part of the GNOME project is to build statistical models for NP form determination by training on annotated sub-corpora. Such statistical models take the form of decision trees (Breiman et al., 1984), which assign probabilities to different NP types according to the input semantic and discoursal features. The correctness of the decision trees relies heavily on the reliability of the annotation. As a result, GNOME places emphasis on achieving reliable annotation, a stage which appears to have been skipped in much other work.

Our work on NP modifiers in GNOME is along the same lines. It aims at finding reliable evidence for which information human authors like to convey in an NP and how they realise it. However, we do not intend to address the first problem in great detail because content selection is usually domain specific. Rules for one domain are not likely to be portable to a different domain since the communicative intentions and information needs can be very different. So our attention is on the second problem, i.e., given a piece of information, how it is realised in an NP. The relevant factors might include NP forms, the semantic properties of a piece of information, discourse properties and communicative goals. Features for NPs are annotated according to (Poesio, 2000b), which include NP types, syntactic features and discourse attributes, etc., (relevant NP features will be introduced when they are needed). We concentrate on modifiers and identify three main features for each NP modifier:

- The pragmatic feature: why is a modifier used in an NP?
- The semantic feature: which property of the entity denoted by an NP is expressed

as a modifier in that NP?

- The realisation feature: which syntactic position is assigned to a given property in an NP? e.g., prenominal or postnominal, adjectival or as a relative clause.

Through training a statistical model on a corpus annotated with the above features, we intend to answer the question of what will be the probability of a given piece of information occupying a given syntactic position on the basis of the semantic and pragmatic properties of that information and relevant NP features, e.g., whether a certain colour attribute should be expressed by means of a prenominal adjective or a prepositional phrase in a definite NP. Notice that it is not possible to use corpus annotation to determine the likelihood of a given property to be chosen, unless we know in advance all of the properties that can be attributed to a given object. Therefore, the model targets the realisation of a property in an NP but not the selection of a property or realisation outside an NP, although some tentative statements can be made about content selection.

To annotate modifiers with the above three features, we need to make clearer distinctions between the possible values of pragmatic and semantic features than we did for the first analysis.

The Pragmatic Feature

We have observed three distinct functions of modifiers in NPs:

1. Providing properties to uniquely identify the concepts denoted by an NP.
2. Having no effect in constraining a unique or unambiguous concept out of an NP, but being important to the situation presented in the main proposition containing the NP.
3. Providing additional details about the referent of a definite or an indefinite NP.

Below we describe the three functions in detail.

1. Providing properties to uniquely identify the objects or concepts denoted by an NP. That is, with such modifiers we can uniquely identify the object/concept or the set of objects denoted by the NP, whereas without them the NP can denote more than one object/concept or sets of objects and therefore is ambiguous in its interpretation. In Example (4.2), the parts in boldface help to uniquely identify the inventory.

   (4.2) *The* ***posthumous*** *inventory* ***of the French king Louis XIV's possessions in 1720*** *describes the table in considerable detail.*

   The identification of such modifiers can be based on two NP features: *logical form type*, which specifies whether an NP is a quantifier, term or predicative, and *genericity*, which specifies whether the object denoted is a generic or specific reference. These have been marked up in the GNOME corpus using the NP feature annotation manual (Poesio, 2000b) and the reliabilities in terms of Kappa statistic (Siegel and Castellan, 1988) are .74 and .82 respectively. The identification can be illustrated by the algorithm in Figure 4.1.

Suppose we have an NP *NP1* and a modifier *M1* inside *NP1*; the NP without *M1* is *NP2*. When one of the following two sets of conditions is satisfied, *M1* is a modifier providing identifying property:

NP1 refers to particular objects, i.e., "unique physical entities, located at a particular place in space or time" (Lyons, 1977),
NP1 intends to identify its referent (indefinites do not identify (Kronfeld, 1990)),
M1 is necessary for *NP1* to uniquely identify its referent.

or

NP1 and *NP2* refer to classes or types,
M1 restricts *NP1* to a subtype of the type denoted by *NP2*.

Figure 4.1: Conditions to be satisfied by modifiers for identification

In addition, when an NP is a definite predicative phrase, it represents a concept that can have only one interpretation (Loebner, 1987). It is often the case that the modifiers inside the NP make it possible to use a definite determiner. The modifiers having such effect include some prepositional phrases, superlatives, or-

dinals and adjectives like *next, last, only, same*, etc., and they should be classified as identifying modifiers. For example,

(4.3) a. *It is the* ***best looking*** *food* ***I ever saw.***

b. *Purple, white and green were the colours* ***of the suffragette movement.***

The modifiers described above subsume those normally considered by the referring expression generation module of an NLG system for uniquely identifying the referents, i.e., components of the referring part of an NP.

2. Having no effect in constraining a unique or unambiguous concept out of an NP, but being important to the situation presented in the main proposition containing the NP in one of the two ways below. The NP is either already unique/unambiguous or not required to have such an interpretation.

Modifiers	Examples
in predicative or quantified NPs	*This is a* ***mighty empty*** *country.* *which seems to argue against any* ***single*** *place* ***of manufacture***
in specific NPs	*Besides, she had a* ***sweet*** *face* ***that attracted him.***

Table 4.3: Examples of semantically important modifiers

- The modifiers express essential pieces of information, because of which the main propositions are produced. Without these modifiers, the main proposition would be redundant. The examples in Table 4.3 illustrate this.
- The modifiers support the situation presented in the main proposition containing the NP in a way other than just providing additional detail about the referent of the NP. For example, the modifiers can give a cause for the volitional action presented in the main proposition or form a succession relationship between themselves and the main proposition. These modifiers are mainly in specific NPs. For instance,

(4.4) *In spite of his* ***French*** *name, Martin Carlin was born in Germany and emigrated to Paris to become an ebeniste.*

In Example (4.4), the modifier *French* takes part in a concession relation between the subordinate phrase and the main proposition, and therefore increases the reader's positive regard for where Martin Carlin was born. If it is removed, the whole proposition would convey a very different meaning or sound strange. Such modifiers will be further discussed in Chapter 5.

These modifiers are more related to the main proposition as a whole rather than just the NP they modify. There is some similarity between NPs containing such modifiers and the *conversationally relevant* descriptions of (Kronfeld, 1990), which exhibit a type of relevance not only to its usefulness for identification but also to its specific context. Such descriptions are called attributive descriptions in (Donnellan, 1977), whose main function is to convey information directly contributing to the communicative goals of a discourse.

3. Providing additional details about the referent of a definite NP which can already uniquely refer to the referent independent of the existence of the modifiers, or of an indefinite NP, which does not intend to identify. Removing these modifiers from an NP would not make the NP ambiguous or affect the situation presented in the main proposition containing this NP in any way, except that less information about the referent is provided. That is, the main difference between an NP with and without the modifiers is the quantity of information being expressed about the referent. Such modifiers are mainly in specific NPs and sometimes in indefinite predicative NPs. For example,

   (4.5) a. *a small house built for the King's mistress,* ***Madame de Montespan,*** *on the grounds of the Palace of Versailles.*

   b. *128 is a bolt-fibula* ***found in the Campagna.***

   The use of such modifiers signifies the presence of an OBJECT-ATTRIBUTE ELABORATION relation between the main proposition and the NP modifiers. They include the modifiers normally generated by an aggregation module, in particular one performing embedding, e.g., (Shaw and McKeown, 1997; Cheng, 1998).

   The effect of such modifiers is usually local to the heads they describe rather than to the main propositions as a whole, which is the main difference between

them and the second type of modifiers. These two types form the non-referring part of an NP.

In terms of their importance to an NP and the main proposition containing the NP, the three functions can be ordered as $1 \prec 2 \prec 3$. It is possible that a modifier demonstrates multiple functions in an NP, in which case we will take the more important one (the one comes first) as its main function. The above classification can be used for the modifier usage in all types of NPs.

The Semantic Feature

Because of the deficiencies of the Upper-Model, i.e., both the GUM concepts and the distinctions between them are not clearly defined, we tried to seek different approaches for classifying the semantics of modifiers. Our method includes two aspects:

- We identify some regular patterns of modifiers in the GNOME corpus and summarise these patterns into semantic categories based on what has been presented in the linguistic literature, e.g., (Levi, 1978; Quirk et al., 1985; Meyer, 1992). The categories are also intended as a refinement of the semantic characterisations of modifying relations in the NIGEL grammar (Mann and Matthiessen, 1985), where correlation between certain semantic properties and the positions of modifiers is proposed. This refinement will allow us to test the correlation on a finer ground. We manually assign these predefined semantic categories to mainly modifiers other than adjectives, which will be illustrated in Section 4.3.2.
- We use WordNet (Fellbaum, 1998) to classify adjectives in order to avoid the ambiguity encountered in using the Upper-Model classification. This choice is also driven by the availability and popularity of WordNet. We introduce this approach in the rest of this section.

In WordNet, the basic semantic relation is synonymy and sets of synonyms (*synsets*) form the basic building blocks. Nouns are organised into hierarchical structures by the class inclusion or subsumption relation (*hyponymy*). Adjectives are loosely divided into two categories: *descriptive adjectives* and *relational adjectives*. A descriptive adjective

typically ascribes a value to a noun concept, e.g., *round* gives a value of *shape*. WordNet contains pointers between descriptive adjectives and the nouns by which appropriate attributes are lexicalised. Descriptive adjectives are organised by the antonymy relation. A pair of words that can form antonyms (e.g., heavy/light) form a *head synset*, around which other synonyms (*satellite synsets*) cluster. A relational adjective is associated semantically and morphologically with a noun (Levi, 1978). It usually resembles a modifying noun and functions as a classifier. For example, *atomic* is a relational adjective and it pertains to the noun *atom*.

As mentioned in Section 3.2.2, we need to map adjectives to concepts in a predicate ontology. Since WordNet has hierarchies for nouns and connections between nouns and adjectives, we can use these to derive the corresponding predicate concepts for adjectives. By using WordNet, mapping an adjective to a predicate concept becomes a three-step process:

1. Choose the correct sense for the adjective (since concepts in WordNet usually have multiple senses);

2. Map the adjective to the noun concept for which the sense ascribes a value, which has two cases:

   - For an adjective in a head synset, there is usually a noun that names the attribute for which the adjective gives a value (directly or through related head synsets). The hierarchy of this noun concept is what we need.
   - For an adjective not in a head synset, find its head synset first and continue as above.

   The noun concept hierarchy contains several concepts related by subsumption. We will explain which concept to choose in Section 4.3.2.

3. Map the chosen noun concept into a predicate concept by appending *-ascription* to it. This predicate concept is subsumed by *property-ascription* in the GUM.

For example, suppose we want to derive the semantic category for *innovative* in the NP *the innovative use of materials*, and the second sense of *innovative* ("being or producing

something like nothing done or experienced or created before") is chosen by a human annotator (represented as *innovative2*). As this word is not in a head synset, we have to find its head adjective first, which turns out to be *original3* in WordNet. *original3* ascribes a value to the noun *originality2*, which is a kind of *quality1* => *attribute2* => *abstraction6*, with increasing abstraction. This is the noun hierarchy derived from WordNet and we choose to stop at the *quality1* level (see Section 4.3.2 for more detail). The corresponding predicate concept is *quality1-ascription*, which is assigned to the semantic category of the adjective *innovative*.

Satisfactory agreement among human subjects on choosing senses for words has been recorded (Fellbaum, 1998) and the mapping to nouns can be done automatically. So this approach is considerably better than manually assigning GUM concepts.

However WordNet has its limitations. There may be phrases and senses that are not recorded. For these modifiers, a predefined semantic category has to be assigned.

Now that we have more refined classifications, we need to know if the distinctions we made can be identified by humans reliably. So we asked human subjects to annotate a part of our corpus with these features to see if they agreed with one another in their annotation. In the next two sections, we describe this process and the results from analysing the annotated corpus.

4.3.2 Annotation Overview

We wrote an annotation manual for the modifiers in NPs, describing which elements of an NP should be marked as modifiers and how to mark their features. XML (eXtensible Markup Language) is used as the markup language, where the start and end of a data field, record or logical group of records are identified by a pair of XML tags with a leading tag delimited by "< ... >" and a trailing tag by "< /... >".

Each modifier is marked with a MOD tag, for example,

> The <mod id="m1"> posthumous </mod> inventory <mod id="m2">
> of the <mod id="m3"> French </mod> king <mod id="m4"> Louis
> XIV's </mod> possessions <mod id="m5"> in 1720 </mod></mod>

We used the same scheme to annotate referring and non-referring modifiers in all types of NPs. Below we briefly introduce the annotation scheme. More details can be found in (Cheng, 1999).

What Is Marked as a MOD

Except for the head and the determiners (excluding possessive ones), all the components of an NP are considered modifying constructions, including adjectives, nouns and noun compounds, prepositional phrases, relative clauses, appositive components and possessive determiners.

One type of construction worth special attention is apposition. The units in appositions are constituents of the same level and they must refer to or describe the same object or else the reference of one must be included in that of the other (Quirk et al., 1985). Appositions have many constructions, but we are most interested in the NP + NP structure. Here are some examples from (Meyer, 1992):

(4.6) a. `My friend John` *is on the phone.*

b. `The first twenty thousand pounds, the original grant,` *is committed.*

c. *The nitrogen in* `organic matter (dead roots and shoots, manure, soil humus, etc.)` *is changed during decomposition to an ammonium form.*

To generate apposition in the same way as other types of NPs, we need to distinguish between the appositive units. Using the terms of (Quirk et al., 1985), in an apposition, one of the appositive units acts as the *defined* expression and the other the *defining* one (called the *definer*). The head of the defined unit is the head of the whole NP and the definer is a modifier to the head.

Generally, we assume that the first unit in apposition is the defined unit (therefore the superordinate phrase) and the second unit the defining unit. This includes all constructions of two different information units separated by a comma, e.g., *the French king, Louis XIV* or *Louis XIV, the French king*, and most constructions of only one information unit, e.g., *the French king Louis XIV*. However, when we have phrases like *financial expert Tom Timber*, the second unit is defined by the first one, which cannot exist on its own. So *Tom Timber* is the head and *financial expert* the modifier.

Attributes of MODs

Each MOD has four main attributes, which are:

1. *ID*: a unique identifier.

2. *TYPE*: the type of a modifier. It is mainly based on the syntactic characterisations of modifiers. The possible values are given in Table 4.4.

`appos`	appositive modifiers.
`such`	modifiers in NPs of the form "NP ***such as NP***", "such NP ***as NP***", "NP ***as NP***" or "NP ***like NP***" (the phrases in boldface are modifiers).
`poss`	possessive determiners.
`preadj`	adjectives before the head.
`prenoun`	modifiers in the form of nouns or noun compounds before the head.
`postprep`	prepositional phrases after the head.
`postpart`	present and past participles after the head noun, which are usually reduced forms of relative clauses.
`postnp`	non-appositive modifiers in the form of noun phrases after the head noun, which are usually the reduced forms of prepositional phrases.
`rel-cls`	relative clauses.
`other`	for all other types not mentioned above.

Table 4.4: Possible values of the *TYPE* feature of modifiers

3. *PRAGM*: the pragmatic feature of a modifier. Corresponding to the three distinct modifier functions identified in Section 4.3.1, the possible values of *PRAGM* are `unique`, `int` and `attr`. The decision follows the algorithm in Figure 4.2.

4. *SEM*: the semantic feature of a modifier. The predefined *SEM* values and examples are given in Table 4.5, where for consistency we attach the WordNet sense number after the corresponding concept.

   There is a tradeoff between the number of values and the achievable agreement on their annotation because the more values a feature has, the less agreement the annotation can expect to achieve. To avoid having too many categories, we use more general concepts to mark the *SEM* feature of an adjective. This decides which noun concept should be chosen in the derived noun hierarchy. The selection of the generalisation levels takes into account the total number of

To decide which value should be assigned to the *PRAGM* feature of a modifier of the NP *NP1*, assume that the NP without the modifier is *NP2*, and the sentences with *NP1* and *NP2* are *S1* and *S2* respectively. We follow the algorithm below:

if *NP1* uniquely or unambiguously refers to some objects or concepts while *NP2* does not
PRAGM = `unique`
else if the meaning of *S2* is incomplete or redundant compared with that of *S1*, or if the meaning of *S2* is dramatically different from that of *S1*
PRAGM = `int`
else *PRAGM* = `attr`

Figure 4.2: The algorithm for annotating the *PRAGM* feature

semantic concepts available for annotation. As we have mentioned in a previous example, the *SEM* feature of *innovative* is marked as `quality1`. Similarly, *round* as in *the round table* would be marked as `spatial-property1`. These general categories mainly include (as defined in WordNet):

`temporal-property1`: a property relating to time. For modifiers like *earlier, final.*

`visual-property1`: attributes of vision, including texture, lightness, colour, etc. For modifiers like *red, dark, superfine.*

`spatial-property1`: any property relating to or occupying space, including dimensionality, shape, form, contour, symmetry, etc. For modifiers like *round, hollow, curved.*

`property2`: a basic or essential attribute shared by all members of a class.

`quality1`: an essential and distinguishing attribute of something or someone.

Note that more specific categories can be refined from `property2` and `quality1`. In fact, the first three categories are subsumed by `property2` in WordNet.

We modified the WordNet window-based browser interface to make it capable of deriving semantic categories of adjectives from the WordNet ontology. The interface is mainly composed of two windows. The upper window shows the word senses retrieved from the WordNet database, and the lower window shows the corpus text with the MOD tags, including all feature-value pairs that have

`location1`	spatial positioning (at a point or in an area), for modifiers indicating where the object denoted by a head is located in physical space. This includes the origin of the object, i.e., the place where the object begins or where it springs into being. *a pattern of brass and pewter* ***on a tortoiseshell ground***
`time-period1`	temporal positioning (in a period of time), for modifiers that indicate the time period the object denoted by the head is located. *the French king Louis XIV's possessions* ***in 1720***
`material1`	cases where the modifier indicates the material of/from which the object denoted by the head is made. *This table's marquetry* ***of ivory and horn***
`identify2`	cases where the modifier names or identifies the referent of the head. The object denoted by the modifier is more specific than that denoted by the head and the two objects are normally (but not necessarily) of the same type. *the practice* ***of veneering furniture with marquetry of tortoiseshell, pewter and brass***
`rephrase1`	cases where the modifier "paraphrases" the lexical content of the head. In this case, the modifier and the head are equally specific and are of the same type. *high blood pressure* ***(hypertension)***
`characterize1`	cases where the modifier provides general "characteristics" of the object denoted by the head. ***Finnish artist*** *Janna Syvanoja*
`content2`	subject matter, with the meaning "on the subject of", "concerning", i.e., the modifier specifies what the head is about. *a book* ***about English grammar***
`subject7` or `object3`	cases in which the modifier occupies the subject or object role of the action denoted by the head. ***the boy's*** *application* (the boy applied for ...) ***the boy's*** *release* (... released the boy)
`purpose2`	the modifier indicates the purpose or function of the objects denoted by the head, i.e., what the objects are used for. *three small drawers* ***for rings***
`possess` or `possinv`	possessive relations between the object(s) O_h denoted by the head and the object(s) O_m denoted by the NP in the modifier in a general sense. The relation can be expressed as O_h *has* O_m (**possess**) or O_m *has* O_h (**possinv**). The relations include the following subtypes: *whole/part, type/instance, set/subset, owner/owned, object/property* and *object/role-relation.* *desks* ***with interiors***, *the name* ***of the maker***

Table 4.5: Predefined semantic categories of modifiers

been annotated. When the annotator chooses a sense for an adjective, the automatically derived predicate concept will be inserted into the annotation for that word.

To mark a *SEM* feature, the annotator should first try the predefined categories in Table 4.5. If none fits, then try WordNet to derive the category automatically, which will be one of the above five values. `other` is used for those whose semantics cannot be obtained from the above means.

We asked two annotators to read the manual, and then mark the NP modifiers in a small corpus according to their understanding of the manual. We measured their agreement on the features being marked and analysed the problems that caused disagreement. Based on this, we revised the manual and trained the same annotators. When the agreement became satisfactory, we asked a trained annotator to mark parts of the GNOME corpus. From the annotated corpus, we could discover regularities in the usage of NP modifiers and design modifier generation algorithms based on these observations.

4.3.3 Results of the Annotation-Based Corpus Analysis

We analysed the annotated museum texts in the GNOME corpus, which contain 1863 modifiers altogether. Our analysis focused on modifiers marked as $PRAGM =$ `attr`, which provide additional information about domain objects.

Agreement on Modifier Annotation

In natural language processing, researchers often use percent agreement between subjects to illustrate the reliability or replicability of their results. This approach as argued by Carletta (1996) is not very revealing because chance agreement is not excluded. She suggests that the Kappa coefficient (K) (Siegel and Castellan, 1988) should be used instead, which "measures pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement". According to (Carletta, 1996), a value of K between .8 and 1 indicates good agreement, and a value between .6 and .8 indicates some agreement.

The agreement on the three modifier features by means of the Kappa statistic is:

Features	Type	Pragm	Sem
Agreement (K)	.97	.77	.81

This demonstrates fairly good agreement on *TYPE* and *SEM* and some agreement on *PRAGM*. The agreement on *PRAGM* shows that the distinctions we are trying to make are relatively clear and human subjects can distinguish between the different uses of NP modifiers to some extent. The main ambiguity exists between `int` and `attr` modifiers. There seems to be a gradual difference between them and where to draw the line is a bit arbitrary. Some disagreement is also caused by the errors in the logical form type and genericity annotation, although good agreements have been achieved on these NP features (Poesio, 2000a).

What is Expressed as a Modifier?

Tables 4.6 and 4.7 show the distributions of the semantic features of modifiers with respect to their functions. In Table 4.6, each cell gives the number of modifiers found in the corpus for each *SEM* and *PRAGM* combination and what percentage such modifiers occupy in those with the same *PRAGM* value. The percentages illustrate the differences in modifier usage. Among the `attr` properties, some appear more frequently than others and they are ordered in decreasing frequencies in the table.

In Table 4.7, each cell gives the number of modifiers found in the corpus for each *SEM* and *PRAGM* combination and what percentage such modifiers occupy in those with the same *SEM* value. It shows that `characterize1`, `spatial-property1`, `visual-property1` and `rephrase1` are more often given as additional information than as other types of information.

Properties such as `possess/possinv`, `location1`, `identify2`, `subject7/object3`, `quality1`, `time-period1`, `material1`, `purpose2`, `temporal-property1`, `state4` and `content2` (in decreasing frequencies, highlighted in Table 4.7) tend to be used more often for referring. This gives a possible order for selecting properties to refer to a discourse entity. It seems to us that the semantic feature itself is far from sufficient for deciding the use of `int`-modifiers, so we do not discuss them here.

SEM	PRAGM		
	attr	unique	int
location1	88 (17.3%)	119 (12.1%)	23 (6.2%)
possess or possinv	59 (11.6%)	239 (24.4%)	58 (15.7%)
identify2	56 (11%)	111 (11.3%)	11 (3%)
material1	45 (8.8%)	52 (5.3%)	13 (3.5%)
other	43 (8.5%)	71 (7.2%)	63 (17%)
time-period1	34 (6.7%)	58 (5.9%)	9 (2.4%)
characterize1	33 (6.5%)	2	0
spatial-property1	32 (6.3%)	7	7
property2	25 (4.9%)	34 (3.5%)	24 (6.5%)
visual-property1	21 (4.1%)	18 (1.8%)	12 (3.2%)
purpose2	15 (3%)	39 (4%)	21 (5.7%)
quality1	14 (2.8%)	76 (7.8%)	77 (20.8%)
state4	10 (1.96%)	16 (1.63%)	8
rephrase1	8 (1.6%)	0	0
subject7 or object3	7 (1.5%)	77 (7.9%)	18 (4.9%)
content2	5	15 (1.6%)	11 (3%)
temporal-property1	3	32 (3.3%)	9 (2.4%)
Total	509 (27.4%)	981 (52.7%)	370 (19.9%)

Table 4.6: The distribution of *SEM* with respect to *PRAGM*

In both tables, We miss out those percentages that are obviously too small (say $<$ 1.5%). Note that the table only lists the main semantic and syntactic categories, so the numbers in a column do not necessarily add up to the amount in *Total*.

Adding `attr` properties can cause confusion sometimes. For example, they might be read as referring information and confuse the reader about the referent (see Chapter 3 for more discussion). Avoiding such ambiguities in generation is important. So we rank *SEM* values with significant preferences for serving `attr` function, e.g., `characterize1`, over those occurring even more frequently as `attr` modifiers, e.g., `location1`.

These observations suggest some preferences for selecting `attr` properties to describe a discourse entity. Below gives an example, where A $\prec$ B means A is preferred over B:

`rephrase1` $\prec$ `characterize1` $\prec$ `spatial-property1` $\prec$ `visual-property1` $\prec$ `location1` $\prec$ `identify2` $\prec$ `material1` $\prec$ `time-period1` $\prec$ `property2`

And the following properties should not normally be chosen or should be down the list: `possess/possinv`, `quality1`, `subject7/object3`, `temporal-property1` and

SEM	PRAGM		
	attr	unique	int
rephrase1	8 (100%)	0	0
characterize1	33 (94.3%)	2 (5.7%)	0
spatial-property1	32 (69.6%)	7 (15.2%)	7 (15.2%)
visual-property1	21 (41.2%)	18 (29.4%)	12 (23.5%)
material1	45 (40.9%)	52 (**47.3%**)	13 (11.8%)
location1	88 (38.3%)	119 (**51.7%**)	23 (10%)
time-period1	34 (33.7%)	58 (**57.4%**)	9 (8.9%)
identify2	56 (31.5%)	111 (**62.4%**)	11 (6.2%)
property2	25 (30.1%)	34 (41%)	24 (29%)
state4	10 (29.4%)	16 (**47.1%**)	8 (23.5%)
other	43 (24.3%)	71 (40.1%)	63 (35.6%)
purpose2	15 (20%)	39 (**52%**)	21 (28%)
possess or possinv	59 (16.4%)	239 (**67.1%**)	58 (16.3%)
content2	5 (16.1%)	15 (**48.4%**)	11 (35.5%)
quality1	14 (8.4%)	76 (**45.5%**)	77 (46.1%)
subject7 or object3	7 (6.9%)	77 (**75.5%**)	18 (17.6%)
temporal-property1	3 (6.8%)	32 (**72.7%**)	9 (20.5%)

Table 4.7: The distribution of *SEM* with respect to *PRAGM*

content2.

Tables 4.8 and 4.9 give the number of modifiers found in the corpus for each *TYPE* and *PRAGM* combination and what percentage such modifiers occupy in those with the same *PRAGM* and *TYPE* values respectively. They show that a syntactic position can be used for any type of modifier. There is a tendency for appositive components, posthead participles and relative clauses to be used more often for realising attr properties, and possessive determiners, prehead adjectives and nouns and prepositional phrases more often for referring properties.

Linguistic work has shown that in human written texts, there are certain preferences about the place where new information usually appears. Fraurud (1990) observes from her corpus that 75% of the complex definite NPs are discourse new references, but her usage of complex NPs mainly refers to those with referring modifiers. In our corpus, we have also found a preference for the place where non-referring information usually appears. 67.19% of such information appears in discourse new references, including bridging descriptions (21.41%, to be introduced in Section 4.4.1), and only 11.79% in discourse old references. The remaining 21% is in predicative phrases. Therefore,

TYPE	PRAGM		
	attr	unique	int
preadj	135 (26.5%)	337 (34.4%)	186 (50.3%)
postprep	98 (19.3%)	293 (29.9%)	114 (30.8%)
appos	77 (15.1%)	49 (5%)	2
postpart	66 (13%)	37 (3.8%)	19 (5.1%)
prenoun	64 (12.6%)	105 (10.7%)	27 (7.3%)
rel-cls	45 (8.8%)	18 (1.8%)	20 (5.4%)
postnp	10 (2%)	13 (1.3%)	1
such	8 (1.6%)	5	2
poss	2	124 (12.6%)	1
Total	509 (27.4%)	981 (52.7%)	370 (19.9%)

Table 4.8: The distribution of *TYPE* with respect to *PRAGM*

TYPE	PRAGM		
	attr	unique	int
appos	77 (60.2%)	49 (38.3%)	2
rel-cls	45 (54.2%)	18 (21.7%)	20 (24.1%)
postpart	66 (54.1%)	37 (30.3%)	19 (15.6%)
such	8 (53.3%)	5 (33.3%)	2
postnp	10 (41.7%)	13 (54.2%)	1
prenoun	64 (32.7%)	105 (53.6%)	27 (13.8%)
preadj	135 (20.5%)	337 (51.2%)	186 (28.3%)
postprep	98 (19.4%)	293 (58%)	114 (22.6%)
poss	2	124 (97.6%)	1

Table 4.9: The distribution of *TYPE* with respect to *PRAGM*

additional properties of a discourse entity are usually given in its first mention.

How to Realise a Property?

We need a more precise correlation between a semantic and a syntactic feature than just a tendency. So we used the *wagon CART building program* (Taylor et al., 1999) developed at the Centre for Speech Technology Research, the University of Edinburgh to train a statistical model for deciding the syntactic form of a property given its semantic and pragmatic features and the necessary NP information. The construction of CART (Classification And Regression Trees) (Breiman et al., 1984) is a common and powerful method for building statistical models from simple feature data. The NP feature used for training is *CAT*, the type of an NP, e.g., proper name, definite NP,

etc., (Poesio, 2000b). We acknowledge that it is more appropriate to use NP features such as definiteness (definite/indefinite) and reference type (refer by name or by class) to train the model rather than using a surface feature like *CAT*, but the GNOME corpus is not annotated with these features and we have to use *CAT* to simulate them.

The program has two parts: *wagon* and *wagon_test* which trains and tests a statistical model on some given samples respectively. Because the size of the annotated corpus is relatively small, we used a cross-validation method. The construction and testing of a model work as follows:

1. Specifying the input and predicted features for *wagon*. In our case, the input features are the semantic and pragmatic features (*SEM* and *PRAGM*) of a piece of information and the type of NP (*CAT*) this information is in, and the predicted feature is the syntactic positions (*TYPE*) that are used to realise this information.

2. Dividing the corpus into two parts, 9/10s for training and 1/10 for testing.

3. Using *wagon* to train a statistical model on the specified part of the GNOME corpus. The result is a decision tree, whose intermediate nodes are questions concerning input features and leaf nodes probability density functions over all possible values of the predicted feature.

4. Using *wagon_test* to test the model on the 1/10 of the annotated corpus, which gives the correct rate of the prediction.

5. Executing steps two to four 10 times to achieve cross-validation and finally calculating the average correct prediction rate. Only the last output tree is recorded and it will be used for further operations.

The above process not only constructs a model for making realisation decisions in embedding, but also tests the accuracy of the model. A fragment of our trained decision tree is given in Figure 4.3, where a leaf node specifies the choice of a realisation (the *TYPE* value with the largest probability), given all the conditions in the non-terminal nodes subsuming the leaf node. However, in the actual output tree, a leaf node contains in addition a list of *(type probability)* pairs for all the values of *TYPE*. So it is possible to

choose another syntactic position with equal or smaller probability to realise a property if so wish. This model has the same function as a set of rules in a rule-based system.

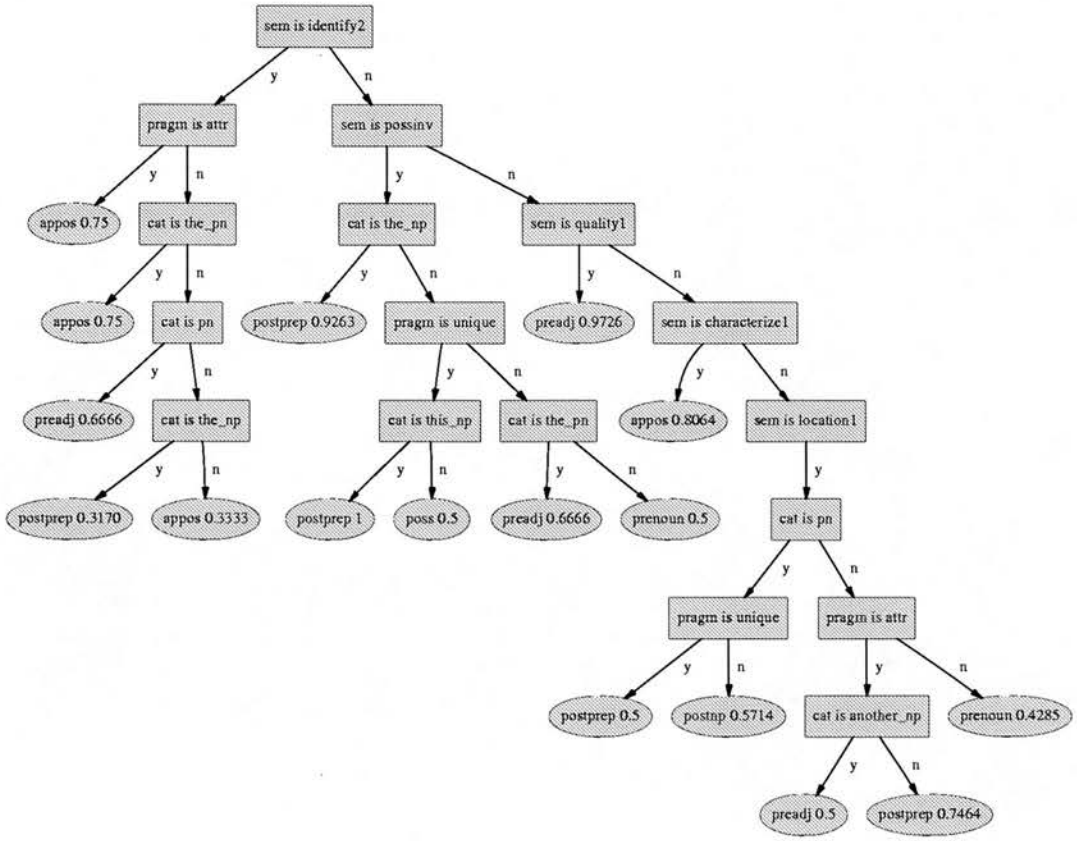


Figure 4.3: A fragment of the decision tree

At the moment, the global success rate for predicting a modifier realisation for museum descriptions is 67.5%. The rates of correct hits with respect to the main values of *TYPE* are shown in the third column of Table 4.10, where *Total* gives the number of modifiers of a specific *TYPE* used for testing. For example, in the annotated testing corpus, there are 124 appositive modifiers. 88.7% of them are predicated correctly by our decision tree. The rates show that the model performs well for appositive, possessive and adjectival modifiers, reasonably well for prepositional phrases and posthead NPs, but rather badly for prehead nouns and posthead participles and the worst for relative clauses.

The low rate is mainly the result of the less refined semantic classification of modifiers. We use `other` for all infrequent properties and information expressed through

relative clauses which cannot be classified to an existing category. About 10% of the *SEM* features of the modifiers in our corpus are given this value. Properties realised as `prenoun`, `postpart` or `rel-cls` are very diverse, so `other` is often used for simplification. There does not seem to be a correlation between specific semantic categories and `rel-cls` or `postpart`. However, the low rate will not be a problem for NLG because relative clause is a catch-all type of realisation and can be used to express all types of information. For stylistic reasons, relative clause is only used in NLG when other forms do not apply. Disregarding `rel-cls` and `postpart`, the global success rate would be 75%. A higher rate might also be achieved by training on a larger annotated corpus.

Type	**Correct Prediction Rates**			
	Museum Descriptions		*Patient Leaflets*	
	Total	Percentage	Total	Percentage
`appos`	124	88.70%	9	77.78%
`poss`	154	88.31%	55	98.18%
`preadj`	570	82.10%	59	96.61%
`postprep`	402	71.64%	44	63.64%
`postnp`	38	63.15%	2	0%
`prenoun`	226	45.13%	35	45.71%
`postpart`	110	25.45%	6	33.33%
`rel-cls`	74	0%	8	0%
overall	1698	67.5%	218	75%

Table 4.10: The accuracy rates of the decision tree with respect to *TYPE*

The decision tree shows that *PRAGM* also plays a role in modifier realisation. This means that there are cases where modifier usage determines or correlates with the syntactic positioning of modifiers. However, it is difficult to single out its effect because of the complex dependencies between *SEM*, *PRAGM* and *CAT*.

Further Testing

The approach we have used allows us to test the realisation model on a different domain as long as there is an annotated corpus for that domain. We annotated a small part of the patient information leaflets in the GNOME corpus. The average success rate of our decision tree in syntactic form prediction in this domain is 75% and its decomposition is given in the last column of Table 4.10. In general, our model is portable to this new

domain. Some degradation in correctness might be due to the small size of the test sample.

Table 4.11 illustrates in detail the behaviour of the decision tree on the test data (some *TYPE* values in the first row are abbreviated to save space). It shows which kinds of mistake the decision tree makes. *Correct* gives the numbers of correct predictions vs. the total numbers, which correspond to the percentages in Table 4.10. For example, for the 44 modifiers annotated as `postprep` in the test corpus, 28 are indeed predicted as prepositional phrases by the decision tree, but 11 are predicted as `preadj`, 4 as `prenoun` and 1 as `appos`. So the correct prediction rate is 63.64%. The table shows that most of the errors are caused by predicting an adjectival form when another form is actually used.

TYPE	**appos**	**such**	**poss**	**adj**	**noun**	**prep**	**part**	**np**	**rel_cls**	*Correct*
appos	7	2	0	0	0	0	0	0	0	7/9
such	0	1	0	0	0	1	0	0	0	1/2
poss	0	0	54	0	0	1	0	0	0	54/55
preadj	0	0	0	57	2	0	0	0	0	57/59
prenoun	0	0	0	10	16	7	0	2	0	16/35
postprep	1	0	0	11	4	28	0	0	0	28/44
postpart	0	1	0	2	1	0	2	0	0	2/6
postnp	0	0	0	1	0	1	0	0	0	0/2
rel_cls	2	0	0	3	1	2	0	0	0	0/8

Table 4.11: Predictions made by the decision tree on the test data

4.3.4 Observations about Proper Names

One goal of the GNOME project is to look at some aspects of NP generation that have received less attention in the NLG literature and improve on the current practice in these aspects. One such aspect is the generation of appositions involving proper names (PNs). Because the corpus is not annotated with information about the inner structures of complex PNs with appositive modifiers, the trained decision tree can only tell us when to use an apposition but not the structure of the apposition.

We had to perform a hand analysis of the proper names in a part of the GNOME corpus. Table 4.12 illustrates the syntactic compositions of the PN related appositions

Syntactic Composition of PNs	Semantic Categories of Definers			Total
	identify2	characterize1	rephrase1	
the-np + PN	the-np/poss-np PN: 15 the-np/poss-np, PN: 5 the-np, known as PN: 1			38.9%
bare-np + PN	bare-np/num-np, PN: 3 bare-np/num-np known as PN: 2	bare-np PN: 4		16.7%
PN + definite-np		PN, PN: 5 PN, the-np: 1	PN (PN): 8	25.9%
PN, indefinite-np		PN, bare-np: 2 PN, num-np: 1 PN, a-np: 2 the-PN, a-np: 2		13%
a-np + PN	a-np, PN: 2 a-np known as PN: 1			5.6%
Total	29(53.7%)	17(31.5%)	8(14.8%)	54

Table 4.12: The syntactic and semantic compositions of appositions involving PNs (the '+' symbol can be substituted by space, comma or "known as")

and their percentages in the collected examples as well as the semantic features of the definers and the corresponding percentages. It shows that in human written texts, there are strong preferences for some constructions over others. While *definite NP+PN* (with or without comma) is the most common construction (about 39% of the analysed complex PNs), *PN+definite NP* is rather rare when the definite NP is not also a PN. We observe the following facts about the complex PN constructions in our corpus:

- Almost all complex PN structures are in discourse new references (52 out of 54).
- Whether a proper name comes first or second in an appositive construction depends on the complexity of the other NP. Table 4.13 illustrates the distribution of modifiers in the analysed appositions consisting of two units. The second unit is generally more complex than the first one. In the examples with complex first units, the units are all indefinite NPs and are separated from the PNs by a comma. So the tendency is: if the common NP is complex (especially with postmodifiers), use the *PN, NP* structure, otherwise use the *NP(,)PN* structure (especially when the NP has zero or only one premodifier). This is consistent with the findings of (Meyer, 1992).
- The use of comma in the *common NP(,)PN* structure depends on the unambigu-

modification	first unit	second unit
No modifier	16	0
One premodifier	12	0
Multiple premodifier only	5	0
Postmodifier	3	6
Pre- and Postmodifier	1	6

Table 4.13: Distribution of modifiers in appositions

ous interpretation of the whole NP. If the NP is ambiguous, i.e., it can denote more than one discourse entity, e.g., *the designer*, no comma should be used as the proper name is vital for the unambiguous interpretation of the NP and it validates the use of a definite determiner. If the NP is itself unambiguous, i.e., it denotes a functional concept in the sense of (Loebner, 1987), e.g., *the queen, the sun*, a comma is used.

- A possessive phrase normally appears as the first unit of an apposition rather than the second to establish a link between the current entity and a discourse old entity, unless the possessive phrase is structurally complex. In our corpus, 7 out of 9 possessive phrases are the first appositive unit and the remaining 2 use *of*-phrase and have premodifiers.

- The use of bare NPs vs. definite NPs in the *common NP PN* structure seems to be arbitrary. We did not find the motivation behind the choice between *the French artist Gilles Jonemann* and *French artist Gilles Jonemann*.

- According to (Meyer, 1992), the *PN, a-np* structure is a bit more common than *a-np, PN* (12% vs. 7% in his corpus), but this is not obvious in our corpus.

Suppose we are realising a discourse entity with a proper name, the above regularities can be summarised into the following algorithm for constructing appositions, which serves as a complement to the decision tree in Section 4.3.3.

- If the entity has a `rephrase1` property, choose *PN (PN)*.

- If the entity has a `characterize1` property, then an apposition consisting of two units, a proper name and a common NP, can be constructed. In this case,

  - If the common NP is complex, e.g., with postmodifiers, choose *PN, common NP*, where the appositive modifier gives a general character of the entity;
  - Otherwise, choose *common NP(,) PN*, where the appositive modifier identifies the entity. The use of comma depends on the unambiguous interpretation of the whole NP.

- The common NP is preferred to be a definite or possessive NP, but a bare NP or indefinite NP is also acceptable.

4.3.5 Summary of Observations from Corpus Analysis

We briefly summarise the results given in Sections 4.3.3 and 4.3.4 here:

- Example preferences for selecting `attr` properties to describe a discourse entity:

  `rephrase1` $\prec$ `characterize1` $\prec$ `spatial-property1` $\prec$ `visual-property1` $\prec$ `location1` $\prec$ `identify2` $\prec$ `material1` $\prec$ `time-period1` $\prec$ `property2`

- The complete decision tree for realising a property given its semantic and pragmatic features and the NP type is shown in Appendix A.2. The tree chooses the syntactic position with the largest probability on a certain feature combination according to the annotated GNOME corpus. The decision about appositions involving proper names is made by the algorithm given on the previous page.

The above rules are more refined and accurate than the first set of embedding rules given in Section 4.2.2. They can be used by NLG systems for the determination and realisation of properties as non-referring NP components. Some of these rules will be incorporated into the implementations to be described in Chapter 7.

4.4 Embedding in Definite Descriptions

In Sections 3.5 and 4.2.2, we have analysed different embedding cases. The situations for non-restrictive clauses, demonstrative phrases and some quality types like evaluative qualities are relatively clear, i.e., they are mostly informing or properties can always

be added to inform. In this section, we will discuss one of the most complex cases of Rule 3.1, embedding descriptive properties in definite descriptions headed by the determiner *the* (we will call them definite descriptions for simplicity). We are interested in the combination of the following two problems:

- Referent properties realisable as adjectives (mainly non-evaluative qualities), pre-head nouns or prepositional phrases can be used for referring or non-referring under different circumstances. When can they be added in a referring expression to provide additional information only?
- When can we embed safely in definite descriptions headed by the determiner *the*? Embedding improperly in such phrases may cause confusion with other objects realised by the same head and make the embedded part sound referring rather than providing new information. Again in Example (3.24), using *the necklace with floral motifs* to describe the current focal object may confuse it with some other necklace which has floral motifs.

Our task is to find out when we can embed some specific types of qualities in definite descriptions safely. Some examples are given in (4.7).

(4.7) a. *The **cylindrical** body of this tsun is divided horizontally into a slightly flared foot, a swelling midsection, and a widely flaring mouth.*

b. *The panelling's **red pine** timber was probably imported from the Baltic, a key trading destination for merchants on the Fife coast.*

Our observation from corpus analysis is that embedding decisions in definite descriptions are related to world knowledge and contextual factors. New information usually appears in discourse new references, in particular bridging descriptions (to be introduced in Section 4.4.1), which usually are definite descriptions. It rarely appears in discourse old definite references.

We use *discourse new definite descriptions* to refer to the first mentions of discourse entities using definite descriptions and *subsequent definite descriptions* to refer to mentions of discourse old objects using definite descriptions, which refer to the same objects as their antecedents and can have the same or different heads. Some examples from

(Vieira, 1997) are: *"the pixie-like clarinetist ... the soft-spoken clarinetist"* and *"a check ... the lost check"*.

Embedding in these descriptions are determined by different factors:

- For bridging descriptions: the degree of obviousness of uniqueness, which is mainly decided by our world knowledge of the relation between the antecedent and the current object, such as whole-part.
- For discourse new/subsequent descriptions: the degree of salience in a certain context, which is affected by a number of factors like whether the object is in the immediate situation.

In the following, we will give heuristics for making embedding decisions considering these two factors.

4.4.1 Embedding in Bridging Definite Descriptions

Bridging descriptions are "the uses of definite descriptions based on previous discourse which require some reasoning in the identification of their textual antecedent" (Vieira, 1997). They have been discussed under different names in the linguistic literature and several types of bridging descriptions have been mentioned, e.g., those identified by Clark (1977).

In descriptive texts, one specific type of bridging description appears more frequently than others. It is called *indirect reference by association* in (Clark, 1977), where a description may not have a directly mentioned antecedent but a closely related one, e.g., *the room ... the ceiling.* This corresponds to the *associative anaphoric use* of the definite article in (Hawkins, 1978) and the *inferrables* in (Prince, 1992). Prince points out that these expressions are special because inferrable entities are technically discourse new, but their existence is assumed to be inferrable by the hearer on the basis of some trigger entities, which are discourse old. Below we use bridging description to refer to this specific type.

The entity that appears in the previous discourse and from which the existence of the current referent (the *associate*) can be inferred (assumed by the speaker) is called the

trigger (Hawkins, 1978). (4.8) gives an example from our corpus, where the phrase in typewriter font refers to the trigger and the underlined phrases to the associates, some of which have embedded descriptive information in boldface.

(4.8) *Situated on a large corner lot,* `this` ***elegant*** `house`*, commissioned by Seldon and Elizabeth Glide Williams, illustrates Morgan's eclecticism. The front facade* ***with its formal symmetry*** *(seven windows across the second register and a central formal entrance), the quoins at the corners, and the frieze around the main door owe allegiances to Renaissance architecture. The* ***iron*** *balcony and* ***Mission tile*** *roof suggest Mediterranean influences.*

This example shows that bridging descriptions with embedded new information appear in human written descriptions. Below we present a heuristic for embedding in such descriptions.

We assume that the knowledge for generating bridging definite descriptions, i.e., relations between domain entities, is present in the knowledge base of an NLG system, although we acknowledge that representing world knowledge is a very difficult problem and few NLG systems have a knowledge base of sufficient coverage and complexity. This is necessary for us to concentrate on the problems we are interested in.

Heuristic 4.1 *Bridging Heuristic: When the referent (R) has a trigger (T) which is mentioned in the previous discourse, and the cardinality of the association relation between the upper-model concepts immediately subsuming T and R is one, embedding will not have a side effect (in the sense described at the beginning of Section 4.4).*

Bridging descriptions can often be represented by an *of*-phrase or a possessive phrase. When the trigger is in focus, the *of*-phrase is often left out. As long as the trigger is uniquely identifiable through saliency or a syntactic construction and the condition in Heuristic 4.1 is satisfied, no confusion will be caused by embedding information because of the obviousness of the relation between the trigger and the associate. Note that the referent does not have to be a singular object. It can be several objects or a set of objects, but in this case, the bridging description has to refer to them as a whole so that the cardinality of the relation remains one.

To test this heuristic, we analysed a small part of the GNOME corpus by hand, which consists of web pages of the Paul Getty Museum. We collected 79 bridging definite descriptions from the Getty texts using the criterion that the definite phrase itself is a discourse new reference but it has a modifier (explicit or implicit) containing a discourse old reference. 55 descriptions have explicit referring components, which are mainly *of*-phrases, possessive determiners and positional prepositional phrases, to denote unambiguous connections between the intended referents and their triggers. 18 descriptions have non-referring components (excluding relative clauses and appositive modifiers) providing additional information about the referents. In these descriptions, the cardinalities of the association relations are all one, and there are no other types of referring components except for *of*-phrases, possessive determiners and positional prepositional phrases. In other words, whenever a bridging description in the analysed corpus carries non-referring information, the cardinality of the relation between the denoted entity and its trigger is one. This small scale corpus analysis partially shows the effectiveness of Heuristic 4.1.

4.4.2 Embedding in Discourse-new/Subsequent Definite Descriptions

For the convenience of discussion, we assume that the referring part of a definite description has been chosen. This assumption does not have to be true in a real generation system. For example, an algorithm like that of Section 7.3.3 can also be used.

The use of a definite description implies that the expression can unambiguously identify the intended referent and there are no potential confusers given all the referring components of the expression. Given this precondition and the salience of the referent, we could decide what to embed. For example,

(4.9) a. *The shape of the* ***jade*** *kuei tablet derived from the stone axe, and was used by the nobility as a symbol of their social status.*

b. *The* ***brand new*** *Research Store, completed in 1993 at a cost of UKP 2.5 million, is Europe's most advanced purpose-built museum store.*

c. *... But this week the polished life of the Duchess of Abercorn took a new turn when she became the focus of a bitter political attack by Sinn Fein. The*

> ***Ulster-based*** *Duchess was prevented from visiting a Catholic primary school in Cookstown after a Sinn Fein councillor claimed parents were threatening a demonstration because she was a member of the British Royal Family.* (From The Daily Mail 22.1.2000)

The *kuei tablet* and the *Research Store* in (4.9a) and (4.9b) are the topics of the two articles respectively, and are both in the visual situation. Therefore they are very salient. In (4.9c), *the Duchess* is the most salient object and new information can be expressed in subsequent mentions.

This phenomenon is also mentioned in linguistic research. In Example (4.10) from (Grosz et al., 1983), the definite appears in the sentence right after the one in which the dog is first introduced and is the most salient entity in that sentence. Since there are no other similar entities in the discourse, no confusion would be produced by adding new information. In fact, conveying additional information, which leads the reader to draw extra inferences, is preferred here over just referring (Grosz et al., 1995).

(4.10) *I took my dog to the vet the other day. The* ***mangy old*** *beast always hates these visits.*

Hence a heuristic for embedding in discourse new/subsequent definite descriptions:

Heuristic 4.2 *Salience Heuristic: When the referent is the most salient object in its context among objects of the same type, embedding would not have a side effect.*

Entities of the same type can be realised by the same head noun, in which case both types of side effect, causing confusion with other objects and making the embedded part sound referring, might happen if additional information is added. We claim that saliency plays the crucial role here.

However, there is no general agreement as to how to measure the salience of an object, which has been proved to be affected by many factors, such as the discourse status of the object and the recency of its last mention, etc. Huls et al. (1995) call these salience related factors *Context Factors (CFs).*

We identify a number of context factors for salience based on previous research in this aspect, mainly (Huls et al., 1995; Gordon et al., 1993):

1. Larger context factors:

   - Topic CF: if the referent is the current discourse topic or in the visual situation (a picture of the object being described is often present in the domain of descriptive text).
   - Recency CF: if the last mention of the referent is the latest among objects of the same type.

2. Smaller context factors (last mention in the previous utterance):

   - Subject referent CF: if the last mention of the referent is the subject of that sentence.
   - Cognitive status CF: if the last mention of the referent is a discourse new reference. This is based on the empirical observation that new information usually appears in the first a few mentions of an entity rather than later ones (Prince, 1992).

If one or more of the above factors are true for a referent, it is likely to be salient. When there is no potential confusers in the immediate context, we can embed attributive information into the corresponding definite description. These factors capture the influence of syntactic prominence, recency and unambiguity on the salience of a referent.

To test Heuristic 4.2, we use the NP annotations in the GNOME corpus (again more details can be found in (Poesio, 2000b)), in particular, the following NP features:

- discourse status: each reference in the corpus is marked with its antecedent information and the connection between the referent and the antecedent if it mentions a discourse old entity or if it is a bridging description. By checking this information, we know if a reference is discourse new, discourse old or bridging.

- `deix`: identifying NPs that refer to objects located in the visual or immediate situation in which the text is being read. If this is the case for an NP, it is marked as `deix-yes`.

- `gf`: marking the grammatical function of an NP in the clause in which it occurs. The values can be `subj`, `obj` and `np-compl` (for the subject, direct object and NP occurring as post-nominal complement of an NP respectively), etc.

The agreements on `deix` and `gf` in terms of Kappa statistic are .81 and .85 respectively. Only a third of our corpus is annotated with these features, and in this we only found 11 definite descriptions with non-referring modifications. Among them, 8 are either `deix=deix-yes` or `gf=subj` or both. The small number of examples is not enough for claiming that the salience factors we give decide the addition of non-referring modifiers in definite descriptions, but there seems to be some correlation between them.

4.4.3 Embedding in Other Types of Referring Expressions

Finally, we briefly talk about embedding in other two types of referring expressions: individual indefinite REs and proper names.

We have mentioned that all modifiers in individual indefinite REs provide additional information about the referent. So there is no further restriction on embedding in these phrases as long as the modifiers do not violate the three rules given in Section 3.5 and `int` and `attr` modifiers are coordinated among themselves.

For proper names, our corpus analysis tells us that new information is mostly expressed in the first mentions in the form of apposition (see Section 4.3.4). There is no difficulty for us to capture these corpus properties and yet satisfy the constraints on syntactic complexity and avoiding side effects.

4.5 Summary

This chapter describes two corpus analyses of NP modifiers in museum descriptions. We aim at finding out which and how additional information is expressed in different types of NPs, in particular referring expressions, in these texts. The decisions are

mainly driven by factors such as the semantic property of a piece of information and how the information is intended to function in the NP.

The first analysis summarises embedding rules using Upper-Model concepts. To overcome the deficiencies of the first analysis, we performed a second analysis using corpus annotation and statistical modelling. We observed some regularities in choosing additional information to be expressed through NP modifiers, and trained a decision tree for determining modifier forms using the annotated corpus. The overall accuracy of the decision tree in predicting modifier forms is 67.5%, but those for `appos`, `poss`, `preadj` and `postprep` are significantly higher. These facilitate a more refined set of embedding rules. Both the method and the results of this analysis can be used by other NLG systems for devising more reliable embedding algorithms.

As to the side effects that might come from embedding, we discuss the most difficult case, embedding non-evaluative information in definite descriptions in a form other than non-restrictive clauses. Factors that might affect this decision are analysed and simple analyses are performed to test our heuristics about them.

Chapter 5

Embedding for Expressing Semantic Relations

It is not a rare phenomenon for human written text to use non-referring NP components to express essential pieces of information or support the situation presented in the main proposition containing the NP. We have shown in Chapter 4 that about 20% of NP modifiers serve such functions. Yet no previous research in NLG investigates this in detail. This chapter briefly discusses the general phenomenon and focuses on the acceptability of using non-referring NP components to express semantic relations that might normally be signalled by "because" and "then" between separate clauses. It describes a psycholinguistic experiment regarding the similarity between the meanings expressed through the above two types of construction. The experiment tests several relevant factors and enables us to accept or reject a number of hypotheses.

5.1 Introduction

5.1.1 `int` Modifiers

We mentioned in Chapter 4 the existence of `int` modifiers, which have no effect in constraining a unique concept out of an NP head, but are important to the situation presented in the proposition containing the NP. For example, the modifier *French* in (5.1) is not for identifying the name, but for establishing a concession relation between

the main proposition and the subordinate phrase to increase the reader's positive regard for where Martin Carlin was born.

(5.1) *In spite of his* ***French*** *name, Martin Carlin was born in Germany and emigrated to Paris to become an ebeniste.*

`int` modifiers can express essential pieces of information, without which the main propositions containing the NPs would be incomplete or redundant. Such modifiers are indispensable to their main propositions. This situation includes many modifiers in indefinite predicative phrases in the GNOME corpus. In previous work on aggregation, predicative phrases are usually places for expressing descriptive information about domain objects, e.g., (Shaw, 1998a). In our domain, about half of the modifiers in indefinite predicative phrases express essential information and half give additional detail.

`int` modifiers can also support the situations presented in the main propositions in a way other than just providing additional information about the NP referents. Such modifiers and their main propositions form the satellites and nuclei of certain argumentative (intentional) and semantic (informational) relations. These modifiers are not indispensable, so removing them will not change the situations presented in the main propositions, but rather make them less convincing.

This second type of `int` modifier is the topic of this chapter because they are not rare in human written text and generating them needs a coordination between text planning and embedding (more motivation will be given in Section 5.2). We are not concerned with the selection of semantic relations, but rather with the questions of whether a relation can be expressed through NP subordination and whether such a realisation conveys similar meaning to a realisation using separate clauses. We leave the discussion about the coordination between text planning and aggregation to Chapter 6.

We have mentioned that NPs with `int` modifiers bear some similarity to the attributive descriptions of (Donnellan, 1977). Donnellan distinguishes an attributive description from a referential description. The former mainly conveys information directly contributing to the communicative goals of a discourse, whereas the latter only enables the

audience to identify a particular referent. (Green et al., 1998) describes an approach for planning attributive descriptions, which represents the two types of description as two distinct types of communicative act in a media-independent plan. Using this compositional, plan-based approach, they are able to decide when and how to select an attributive description to satisfy the higher level communicative goals. This approach is mainly concerned with the high level planning of a description containing information for both identifying the referent and satisfying communicative goals, whereas we are more interested in the addition of information which serves certain goals but is not for identifying. So this chapter discusses a complementary issue to that in (Green et al., 1998).

The results of this chapter mainly serve as a theoretical contribution and they are not fully implemented in ILEX-TS and GA-plan. One reason is that the museum domain offers little variation in the use of semantic relations, both within and between clauses. The other reason is that the implemented systems can only produce limited types of NP and sentence structure, therefore the difference in algorithms cannot be demonstrated through the system output.

5.1.2 Expressing Semantic Relations

It is generally agreed that a rhetorical relation can be realised through different syntactic constructions. In the NLG community, there has been a large amount of research on using different cue phrases and clause orders to realise the same rhetorical relation, e.g., (Knott, 1996; Marcu, 1997b). It is also observed that when a relation is inferrable, the connection word can be left out. However, this previous research focuses on realisations using two separate clauses.

The semantic roles of non-restrictive (NR) NP components, in particular non-restrictive clauses, are mentioned in many grammar and linguistic books. Quirk et al. (1985) point out that an NR clause in a referring expression is usually neutral in its semantic role (i.e., it provides descriptive information about its head), but sometimes it can contribute to the semantics of the main clause in a variety of ways. They summarise three types of semantic relations that can be expressed by an NR clause:

Causal, where the situation in the main clause is caused by that in the NR clause, for example, *He sent ahead the sergeant, who was the most experienced scout in the company.*

Temporal, where the two clauses form a time sequence, for example, *In 1960 he came to London, where he has lived ever since.*

Circumstantial, where the NR clause sets a temporal or spatial framework for interpreting the main clause, for example, *The boy, who had his satchel trailing behind him, ran past.*

Halliday (1985) mentions that a subordinate clause can elaborate a part of its primary clause through restating, clarifying, refining or adding a descriptive attribute or comment, for example, *Inflation, which was necessary for the system, became also lethal.*

Halliday's notion of elaboration is much more general than that in other coherence theories like RST. To avoid difficulties in choosing among facts for text structuring, NLG systems would not normally take the above examples as ELABORATION.

These non-restrictive modifiers are a part of the non-referring part and we use NR for both non-restrictive and non-referring below. For the convenience of discussion, we define some terminology to be used throughout the rest of the thesis:

An NR construction/sentence: a sentence that has a main clause and a subordinate NR modifier attached to one of its NPs, e.g., *Private Eye, which couldn't afford the libel payment, had been threatened with closure.*

A hypotactic construction/sentence: a sentence that has a main clause and a dependent clause, connected by a cue phrase. This is a common way of expressing semantic relations such as causality, for example, *Private Eye had been threatened with closure because it couldn't afford the libel payment.*

An elaboration realisation: a type of hypotactic construction where one clause elaborates the semantics of the other. We take cue phrases *as for* or *what is more*

to signal elaboration relations. We acknowledge that these cue phrases are controversial in their semantic interpretations, but not using cue phrases would be even more ambiguous. Besides, our discussion does not heavily depend on these cue phrases. For example, *Private Eye had been threatened with closure. As for Private Eye, it couldn't afford the libel payment.*

This chapter is only concerned with semantic (informational) relations other than OBJECT-ATTRIBUTE ELABORATION, for example, causal relations, which are normally expressed by a hypotactic construction using cue phrases such as *because*. Argumentative relations are beyond the scope of this thesis.

The above discussion is not completely in line with the heuristics of (Scott and de Souza, 1990). Based on cognitive modelling of human understanding of text, they give heuristics which prefer to "always generate accurate and unambiguous textual markers of the rhetorical relations that hold between the propositions of the message" and require that "embedding can only be applied to the ELABORATION relation". Their heuristics give general guidelines for NLG, but sometimes they can be overridden by stylistic considerations, on which occasions alternative realisations will be preferred.

5.2 Motivation

The above linguistic research suggests for an NLG system the possibility to express certain semantic relations through NR constructions. This is important in two aspects. Firstly, to produce natural text, an NLG system has to choose among possible paraphrases one that satisfies the highest number of constraints in a certain context. An NR construction might give a more concise alternative realisation for a relation, where the relation is expressed implicitly rather than explicitly and usually more subtly. It does not need cue phrases in most cases, and therefore could avoid using cues too heavily. This could be a better realisation under certain circumstances.

Secondly, a major task of text planning is to select interesting relations to structure a text. The decision can be affected by a variety of factors such as the availability of realisation options. For a relation preferred to be expressed implicitly, if the corresponding realisation is not available, the relation cannot be chosen. So an NR

construction enables a wider range of relations (especially those that are preferred to be expressed implicitly) to be selected for text structuring because the corresponding syntactic option is available.

Previous research in NLG mainly focuses on using NR constructions to realise ELABORATION relations but not other semantic relations, e.g., (Scott and de Souza, 1990; Hovy, 1993). The NR component usually adds a descriptive attribute to the object denoted by the NP head. `int` modifiers are largely ignored in previous work on NP generation. Because of their role in supporting the semantics of the main propositions, the selection of `int` properties is a concern of the text planning process to serve the overall goals for producing the text. However, compared with `unique` modifiers, they are less essential for an NP and can only be added if there are available syntactic slots. In this respect, they resemble `attr` modifiers. The realisation of `int` properties is a part of realising non-referring NP properties and should be a task of embedding. Therefore, the generation of `int` modifiers needs the coordination and interaction between text planning and embedding.

To understand how to enable the embedding process of an NLG system to generate such modifiers, we are faced with two questions, which are not answered by linguistic research:

1. Can this type of modifier be identified by human subjects? Or more generally, if an NLG system produces NP modifiers for different purposes (as specified by the values of the *PRAGM* feature), will they be properly understood by humans?

2. Under what circumstances can an NR construction be used in substitution of a hypotactic construction without changing the meaning dramatically and how close are the meanings conveyed by the two representations?

An NLG system must come up with some solutions, simple or complex, to these two questions in order to choose among paraphrases. The first question has been answered positively in Section 4.3 through the agreement found on annotating the *PRAGM* feature of a modifier. The second question is to be answered partially in this chapter. It motivates the empirical experiment described in Section 5.3, which aims at finding

out the factors related to the generation of this type of NP modifier. In particular, it is designed to find out partial answers to the following questions:

- Which rhetorical relations can be expressed through an NR construction?
- How similar are the meanings expressed by a hypotactic construction and the corresponding NR construction?
- For each relation, are there some cases that can be re-expressed through NR constructions, while others cannot? What contributes to the difference within a relation?

`int` modifiers are non-referring NP components, but sometimes it is difficult to tell from the surface of a text whether a component is referring or not. Adjectives and prepositional phrases normally appear the same even when they function differently, so their functions in NPs might be controversial. To avoid disagreement, we restrict our experiment to non-restrictive relative clauses within referring expressions. They can usually be clearly identified, e.g., by punctuation, and are frequently used in human written texts. However, we would expect no substantial difference between a non-restrictive clause and other non-referring components.

5.3 The Experiment – a Detailed Description

Due to the large number of semantic relations and the lack of general agreement on relation sets, it is impossible to conduct a full scale study on all relations. In addition, it is often possible to apply multiple relations to the connection between discourse elements. Therefore, we use cue phrases to signal semantic relations, and reduce the size of the experiment by focusing on only two relations: a causal relation signalled by *because* and a temporal relation signalled by *then*.

The reason for choosing these relations is that the possibilities of expressing them through NR constructions have already been shown by linguists. The two cue phrases are typical for the corresponding relations and they can often substitute other cue phrases for the same relations. Although they might have cross relation usage, we only use one specific meaning here.

We acknowledge that using only one cue phrase for a relation is far from enough for making general claims about that relation. However, we could still get interesting and reliable results, and hopefully this could be the first step toward a more ambitious goal. In the rest of this thesis, we will use the term causal or temporal relation to refer to the specific relation signalled by *because* or *then*.

5.3.1 Independent Variables and Hypotheses

We aim at finding out what determines if a given semantic relation can be expressed through an NR-construction or not. From the generation point of view, our question is: given two facts and the semantic relation between them, what extra input do we need for making realisation decisions?

Since the museum domain offers little variation in the use of semantic relations (normally just OBJECT-ATTRIBUTE ELABORATION), we chose test samples from the Wall Street Journal source data, which consist of descriptive texts in the commercial domain which contain many interesting relations. We collected examples of *because* sentences, and transfered them to NR sentences by hand. Comparing the two constructions, we found some interesting variation. For example, comparing the sentences in (5.2) and (5.3), we found intuitively that the meanings of (5.2a) and (5.2b) are much closer than those of (5.3a) and (5.3b). In other words, (5.2b) can be used in substitution of (5.2a), whereas (5.3b) cannot so easily substitute (5.3a).

(5.2) a. *Private Eye had been threatened with closure because it couldn't afford the libel payment.*

b. *Private Eye, which couldn't afford the libel payment, had been threatened with closure.*

(5.3) a. But P&G contends the new Cheer is a unique formula that also offers an ingredient that prevents colors from fading. *And retailers are expected to embrace the product, because it will take up less shelf space.*

b. *And retailers are expected to embrace the product, which will take up less shelf space.*

(5.4) and (5.5) give another example, where the sentences in (5.4) are much closer in

meaning than those in (5.5). A similar pattern was observed in a number of other collected sentences.

(5.4) a. *The girl decided to leave because she was upset by the activities of the ghost.*

b. *The girl, who was upset by the activities of the ghost, decided to leave.*

(5.5) a. *The banner is one of the Museum's most treasured objects because it was made in about 1520.*

b. *The banner, which was made in about 1520, is one of the Museum's most treasured objects.*

We claim that it is the degree of inferrability of the relation between the semantics expressed through the two clauses that makes the difference. We define the *inferrability* of a causal/temporal relation as:

Definition 5.1 *Given two separate facts, the likelihood of human subjects inferring from their world knowledge that a causal/temporal connection between the facts might plausibly exist.*

In examples (5.2) and (5.3), the fact that Private Eye cannot afford the libel payment is very likely to directly cause the closure threat, whereas a product occupying less space is not usually a cause of it being accepted by retailers according to common sense. Therefore, the two realisations in (5.2) can be used in substitution of one another whereas those in (5.3) cannot.

Inferrability is dynamic and user dependent. Given two facts, people with different background knowledge can infer the relation between them with different ease. If a relation is easily recognisable according to general world knowledge, we say that the inferrability of the relation is *globally strong*, in which case we hypothesise that a hypotactic and an NR construction can express the relation almost equally well (if not considering rhetorical effect). Context can also contribute to the inferrability of a relation. A relation not easily recognisable from world knowledge may be identified by a reader with ease as the discourse proceeds. In this case, we say that the inferrability of the relation is *locally strong*, and we hypothesise that the two constructions can

Independent Variables	Levels	
Relation	causal	temporal
Inferrability	strong	weak
Position	initial	final
Order	hypotactic vs. NR	NR vs. hypotactic
Subordination	nuc subordination	sat subordination
Cued/NoCue	use cue	not use cue

Table 5.1: Independent variables and their values

express the relation equally well only in a certain context. In this chapter, we mainly consider whether a relation is globally strong or not and we will describe how we decide the value of the inferrability of a given relation in Section 5.3.3.

In Table 5.1, we summarise the factors (independent variables) that might play a role in the closeness judgement between the semantics of a hypotactic construction and an NR construction. The levels are possible values of these factors. Besides *Relation* and *Inferrability*, we have the following factors:

Position gives the location of the NP that contains the NR component. It can be the first (`initial`) or the last (`final`) phrase in a sentence (we restrict ourselves to sentences with two top-level NPs in the experiment);

Order gives the order of presentation, a hypotactic sentence to be compared with an NR sentence or vice versa, which is used to balance the influence of cue phrases on human judgement;

Subordination specifies whether the nucleus or the satellite is realised as an NR clause. We assume that in the causal relation, the clause bearing *because* is always the satellite. Since the temporal relation is a multi-nuclear relation, this factor does not apply;

Cued/NoCue means using a cue phrase in the NR clause or not, which is only applicable to the temporal relation, for example,

(5.6) *The health-care services announced the spinoff plan last January, which was then revised in May.*

We are not really concerned with how *Order* affect human judgement, but we have to balance it in the experiment to prevent it from biasing the observation. However if it does have an impact, we are interested to see what it is.

Based on the factors just defined and our observation of human written sentences, we can make hypotheses about the similarities in meaning between the two types of syntactic construction. These hypotheses will be tested through experiment and if accepted, they can be realised in an NLG system.

Suppose we have two facts and a semantic relation between them,

Hypothesis 5.1 *For both causal and temporal relations, the inferrability of the relation between the two facts contributes significantly to the semantic similarities between a hypotactic construction and an NR construction expressing the relation.*

In other words, if the *inferrability* of the relation is strong, the relation can be expressed similarly through an NR construction, otherwise, the similarity is significantly reduced.

Since an NR component usually just elaborates its head, a related question at this point is how *inferrability* of a causal or temporal relation affects the similarity between an NR construction and an elaboration realisation, which leads to another hypothesis:

Hypothesis 5.2 *For both relations, the inferrability of the relation between the two facts contributes negatively to the semantic similarities between an NR construction and an elaboration realisation.*

In examples (5.2) and (5.3), since the inferrability of the relation between the two clauses in (5.2a) is stronger than that in (5.3a), (5.2b) is less similar to its corresponding elaboration realisation than (5.3b) is according to Hypothesis 5.2.

Hypothesis 5.3 *For the causal relation, the satellite subordination bears significantly higher similarity in meaning to the hypotactic construction than the nucleus subordination does.*

For example, (5.2b) would be preferred to "*Private Eye, which had been threatened*

with closure, couldn't afford the libel payment." A similar heuristic is also given in (Scott and de Souza, 1990).

Hypothesis 5.4 *For the temporal relation, both the position of subordination and the use of an appropriate cue phrase in the NR clause contribute significantly to the semantic similarities between an NR construction and a hypotactic construction.*

According to this hypothesis, Example (5.6) would be preferred to the realisation that does not have *then*.

Besides the independent variables in Table 5.1, other factors might also affect human judgement of text. We have to control these irrelevant factors to prevent them from biasing our observation, and make sure that only the factors we specify play a role in the final analysis. Our approach will be described in Section 5.3.3.

5.3.2 The Design of the Experiment

In order to assess the semantic similarity between the two types of construction, which is thought to be influenced by the independent variables, we need to have human judgements on the following two dependent variables:

Naturalness: how fluent a sentence is on its own.
Similarity: how similar the meanings of two sentences are without considering their naturalness.

This separation is to prevent the influence of an unnatural realisation on similarity judgement. The scales of the variables are selected such that all values on the scale have natural verbal descriptions that could be grasped easily by human subjects (see Table 5.2). Similar rating methods have been described in (Jordan et al., 1993) to compare the output of a machine translation system with that of expert humans. Through analysing human assessment of the dependent variables, we hope to accept or reject the hypotheses we made.

In the NLG community, people usually use simple methods, such as comparing the means of scores, to decide if one model is preferred by human subjects over another.

Values	Dependent Variables	
	Similarity	Naturalness
6	exactly the same	N/A
5	very similar	natural
4	more similar than different	fairly natural
3	more different than similar	so-so
2	very different	fairly unnatural
1	totally different	unnatural

Table 5.2: Dependent variables and their values

We wish to use a more reliable method, which would decide the least number of data points to be used and the distribution of the properties we want to test among them.

Since we want to measure different groups of *similarity* judgement based on different *inferrability, order* or *position* levels, a between-groups (independent or randomised) design (Hatch and Lazaraton, 1991) seems to be most appropriate. Therefore the question behind Hypothesis 5.1 is: are the means of the ratings for the two similarity groups with different inferrabilities significantly different? The null hypothesis says that the means are equal. We are to find out whether we accept this or not and if there are other important factors we have not expected.

The between-groups design we use is illustrated in Table 5.3, where all possible combinations of the independent variables are listed. When considering any factor, there is an equal number of data points in each group and the groups are properly balanced for other factors. In the table, *Paraphrases* gives the types of alternative sentence each original sentence has. For example, when the original sentence is an NR construction, we want to know if *inferrability* affects its similarity to a causal and an elaboration hypotactic sentence, hence the causal and elaboration paraphrases. The paraphrases should be scored by human subjects for their similarities to the original sentences and their naturalness.

The selection of the test sample should conform to this design, that is, the test sample should instantiate all of the combinations and contain an equal number of data points for each combination.

Independent Variables				Paraphrases
Relation	Order	Inferrability	Position	
causal	hypotactic vs. NR sentence	strong	initial	nuc & sat subordination
			final	NR sentence
		weak	initial	nuc & sat subordination
			final	NR sentence
	NR sentence vs. hypotactic	strong	initial	causal & elaboration hypotactic
			final	
		weak	initial	
			final	
temporal	hypotactic vs. NR sentence	strong	initial	cued & not cued NR sentence
			final	
		weak	initial	
			final	
	NR sentence vs. hypotactic	strong	initial	temporal & elaboration hypotactic
			final	
		weak	initial	
			final	

Table 5.3: A between-groups design

5.3.3 Collecting the Test Sample

To collect the data, we use a method similar to *random selection* to create a stratified random sample. The sample should contain 12 hypotactic sentences and 12 NR sentences, two for each combination of the causal relation and one for each combination of the temporal relation. These numbers are used to obtain as large a sample as possible which could still be judged by human subjects in a relatively short period of time (say less than 30 minutes). If the sample size is too big, the subject might lose concentration in the middle, which could lead to poor results.

We have mentioned that we use cue phrases as the indicators of the semantic relations between clauses. From the Wall Street Journal source data, we collected all the sentences that contained *because* or *then*, and went through each of them to pick out those that actually signalled the desired relations and could potentially have NR-realisations, i.e., where there was a coreference relation between two NPs in the two clauses. This formed two large sets of hypotactic sentences.

Since the similarities between representations are to be tested from both directions, hypotactic to NR construction and vice versa, sentences containing NR clauses signalled by ", *which*" or ", *who*" were collected similarly. This resulted in a large set of NR

sentences.

We wanted 8 sentences from the *because* set, 4 from the *then* set and 12 from the NR set half of which do signal one of the two relations. So we randomly selected a sentence by category. If it realised an unused factor combination, it was kept in the sample. This process was repeated until we collected the right number of data points which instantiated all combinations of factors in Table 5.3.

When choosing the test sample, we tried to control the syntactic complexity of the sentences to prevent it from biasing the observation. It is obvious that the more complex a sentence is, the more difficult it is to understand its meaning and make an assessment. Since judging the complexity is a hard problem itself, we adopted some simple control methods. We avoided selecting too complicated sentences, for example, those taking more than two lines, but when there was no better alternative, we simplified the sentences by removing some phrases of minor importance to our problem, e.g., embedded verbal phrases.

In addition, some subjects found the business genre hard to understand, but this did not seem to prevent them from making judgements. So we ignored the influence of genre.

We used our own intuition in selecting data bearing a range of different inferrabilities. To see if this intuition is replicable, we asked one subject to judge the inferrability of the same data set according to the questionnaire given in Appendix B.3, which uses Definition 5.1 and a five-point scale for inferrability: 5 for `very likely`, 4 for `quite likely`, 3 for `possibly`, 2 for `even less possibly` and 1 for `unknown`. We took values of 4 and 5 as `strong` and the others as `weak`. Then we calculated the kappa coefficient (K) between the two judgements using the formula below:

$$K = (\text{Observed agreement} - \text{Chance agreement})/(1 - \text{Chance agreement})$$

The K is .67, which allows only a tentative statement to be made. That is, different subjects share their intuition on the inferrability of a relation only to some extent. The author's version was used for the experiment.

For the 24 data points, we manually produced the corresponding paraphrases. These

were then put into a questionnaire in random order for human assessment of the two dependent variables for each paraphrase. The original sentences were given in their context (reduced to a few sentences) to help the subjects to get a better understanding of the relations expressed in the sentences and also make it easier for them to make decisions.

The questionnaire is given in Appendix B.1. Here is an example taken from it, where the original sentence is given in boldface and the context of the original sentence in smaller font. The subjects were asked to rate the similarities of the two alternative sentences to the original one and their naturalness.

1. Mr. McGovern, 63, had been under intense pressure from the board to boost Campbell's mediocre performance to the level of other food companies. **The board is dominated by the heirs of the late John T. Dorrance Jr., who controlled about 58% of Campbell's stock when he died in April.**

   (a) The board is dominated by the heirs of the late John T. Dorrance Jr. because he controlled about 58% of Campbell's stock when he died in April.

   Similarity : 6 5 4 3 2 1

   Naturalness: 5 4 3 2 1

   (b) The board is dominated by the heirs of the late John T. Dorrance Jr. As for Mr. Dorrance, he controlled about 58% of Campbell's stock when he died in April.

   Similarity : 6 5 4 3 2 1

   Naturalness: 5 4 3 2 1

We had ten native English speakers fill in the questionnaire. They ranged from first year undergraduates to Ph.D. students. They were asked to circle the numbers that express their assessment of the similarity between a paraphrase and its original sentence and the naturalness of the paraphrase. Each subject scored all data points, so that the differences in human judgement are within groups rather than between. All together, we have 240 similarity scores and 240 naturalness scores, among which 160 are for *because* and 80 for *then*.

5.3.4 Results and Discussion

This section discusses the results of the experiment and the conclusions that can be drawn from them.

Similarity

Since the similarity data is ordinal data and departs significantly from a theoretical normal distribution according to the One-Sample Komogorov-Smirnov Test, we choose the Mann Whitney U on this data, which is a test for comparing two groups on the basis of their ranks above and below the median. The result is summarised in Table 5.4, with statistically significant items in boldface (taking the conventional p level .05). The Z scores usually tell how many standard deviations above or below the mean an observation might be. They are calculated using different formulas in different tests.

Relation	DependVar	Factors	Z	2-tailed P
causal (160 cases)	Similarity	Inferrability	-4.1015	**<.0005**
		Order	-2.6400	**.0083**
		Position	-.2136	.8308
temporal (80 cases)	Similarity (cued)	Inferrability	-.1022	.9086
		Order	-1.1756	.2398
		Position	-2.0649	**.0389**

Table 5.4: The output of Mann Whitney U on the similarity data

For the causal relation, there is a significant difference in the ranks assigned to the similarities of the two groups of different inferrabilities (P<.0005). This means that more sentences from the group of strong inferrability are ranked high in their similarities to the original sentences. So we have high confidence to accept part of Hypothesis 5.1, that is, in the test sample the strong inferrability of the causal relation between two facts makes the semantic similarities between a hypotactic construction and an NR construction significantly higher than the weak case does. But this procedure can only be used for descriptive, rather than inferential, purposes. In the strong case, around 70% of the paraphrases are given `very similar` or `exactly the same`.

We treated *order* as a factor to be balanced and did not expect it to have a significant

effect, but it does (P=.008). As in Figure 5.1, an NR paraphrase shows much higher similarity to its corresponding hypotactic sentence (the two dark columns) than the other way round (the two grey columns), but the difference becomes smaller for the strong inferrability case. This could be because the causal relations expressed in NR sentences generally sound weaker than those in hypotactic sentences and the cue phrase has a big influence on the perception of a relation.

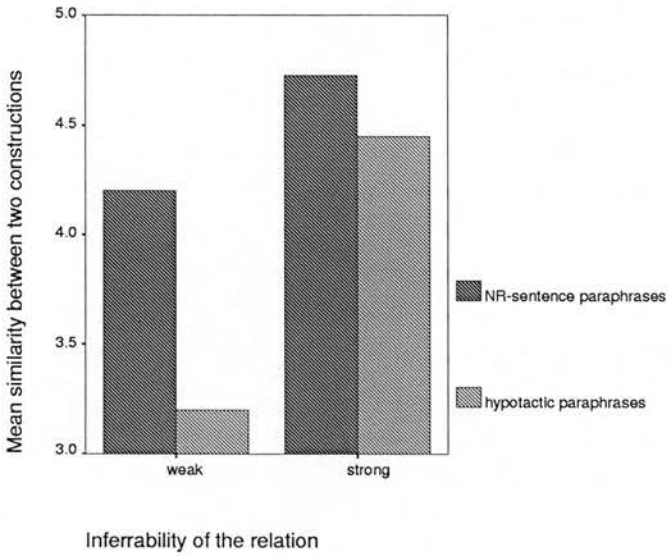


Figure 5.1: The interaction between *order* and *inferrability*

There is a disordinal interaction between *inferrability* and *position* (Figure 5.2), which means that a final position makes a weak case more similar, but a strong case less similar. This might imply that it is preferred for the cause to be given before the result, whereas descriptive information at the end of a sentence.

For the temporal relation, *position* is the only significant factor. So part of Hypothesis 5.4 is confirmed, that is, the final position subordination makes an NR paraphrase significantly more similar to the corresponding hypotactic construction than the initial position does.

We do not have enough evidence to accept the claim that the inferrability of the temporal relation contributes significantly to the similarity judgement (as in Hypothesis 5.1). However, when we graph the similarity scores for the alternative sentences using cue phrases, strong or weak in inferrability, we get 78% `very similar` or `exactly the`

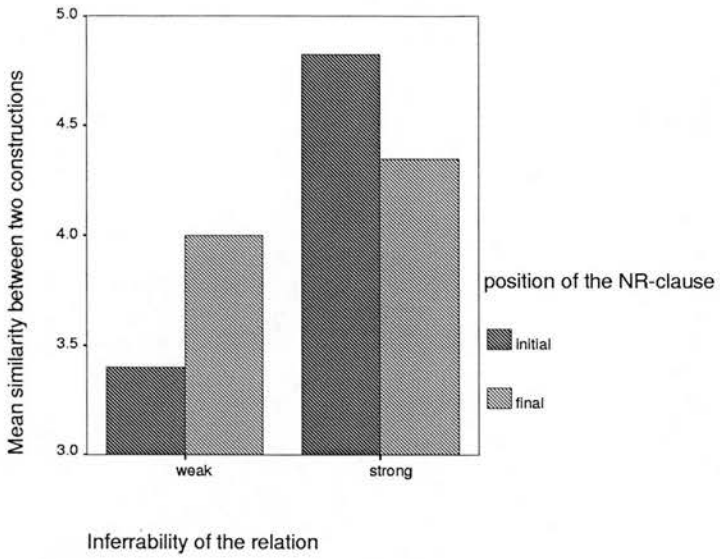


Figure 5.2: The interaction between *inferrability* and *position*

same. Comparing this with that of the strong causal case using the Mann Whitney U test, we get a significance level of 0.0294. This means that we have strong confidence to believe that the similarity rank for the temporal relation if using a cue phrase is significantly higher than that for the strong causal relation. Therefore, the temporal relation can always be realised by an NR construction as long as an appropriate cue phrase is used in the NR clause.

The assumption of normality is also not met by the subset of the data related to Hypothesis 5.3 and 5.4 (i.e., the similarity scores for nucleus/satellite subordination paraphrases and cued/nocue paraphrases). We use the Wilcoxon Matched-Pairs Signed-Ranks Test because we are comparing pairs of paraphrases. The result is given in Table 5.5. We accept the hypothesis that the similarity ranks of nucleus and satellite subordination are significantly different in the initial position (Hypothesis 5.3). This confirms the linguistic observation that information of greater importance should be presented in a main position rather than a subordinate position. We can also accept the hypothesis that for the temporal relation, using cue phrases in NR clauses can significantly improve the similarity score of the NR construction (Hypothesis 5.4).

For the temporal relation, we do not find a significant connection between the simi-

Relation	Paired Variables	Cases	Z value	2-tail Sig
causal	Nuc-similarity/Similarity	40	-3.4954	.0005
temporal	NoCue/Cued	80	-3.02	.003

Table 5.5: The output of the Wilcoxon Matched-Pairs Signed-Ranks Test

larity of the elaboration realisation and *inferrability* using the Mann Whitney U test (Hypothesis 5.2), but for the causal relation, the significance level is .0542. Although this might not be enough for rejecting the null hypothesis, it shows a strong trend that the NR construction of the causal relation with weak inferrability is more similar to the elaboration realisation than the strong case is. This could also mean that an NR component still sounds like elaboration no matter how strong the inferrability is.

Naturalness

We use the Mann Whitney U test on *naturalness* with regards to *order*, *inferrability* and *position*, and find no significant connection. Figures 5.3 and 5.4 show the distribution of naturalness assessment of the paraphrases for the causal and temporal relation respectively. The majority of the NR constructions are `natural` or `fairly natural`, which suggests that they could be good alternative realisations.

There is a significant difference between the naturalness of the elaboration paraphrase and that of the hypotactic paraphrase for both relations ($P < .0005$). In general, the subjects do not prefer the elaboration realisation.

Summary and Further Discussion

We briefly summarise the heuristics drawn from the experiment for expressing the causal and temporal relations with an NR construction. This is an acceptable realisation in the following circumstances:

- The causal relation holds between two facts and the inferrability of the relation is strong, in which case satellite subordination should be used and the cause is preferred to be given first; or
- The temporal relation holds between two facts, in which case a final position sub-

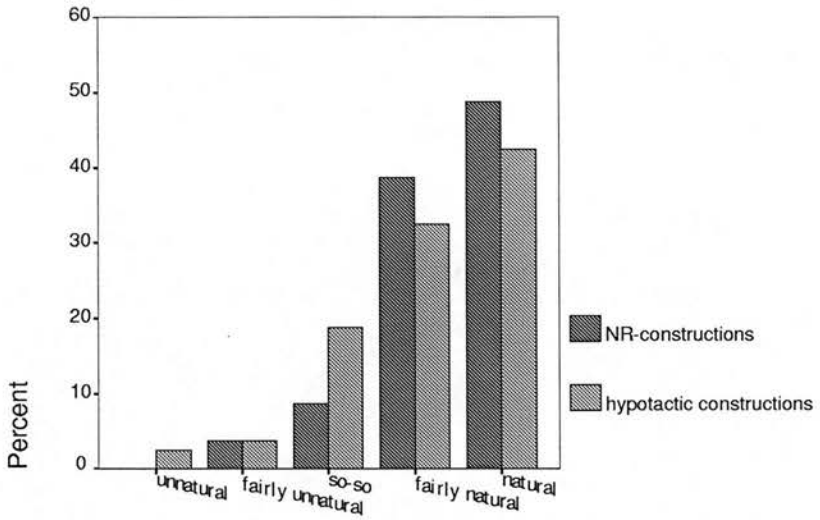


Figure 5.3: The naturalness of the causal paraphrases

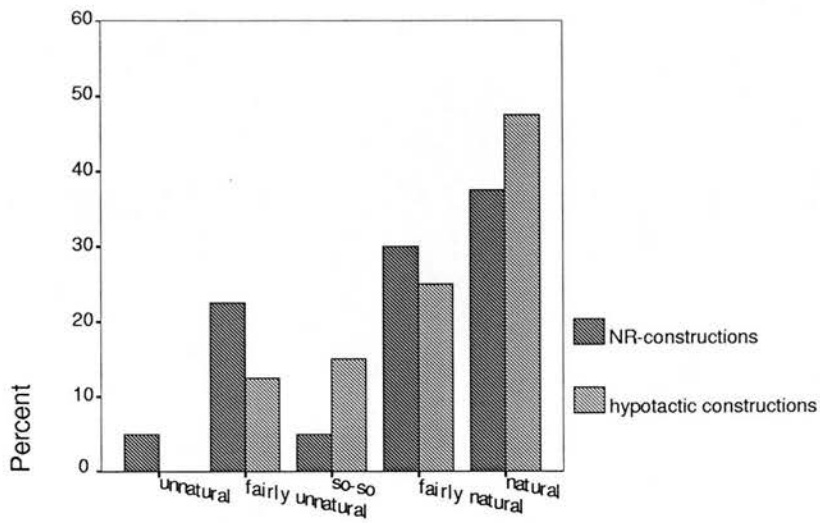


Figure 5.4: The naturalness of the temporal paraphrases

ordination should be used and an appropriate cue phrase, like *then*, is preferred in the NR clause.

These are conservative heuristics for safe generation. We acknowledge that other property combinations might also lead to good realisations in certain circumstances.

We have mentioned that an NR construction can express the causal/temporal relation and the OBJECT-ATTRIBUTE ELABORATION relation at the same time, irrespective of the inferrability of the relation. Generally speaking, a semantic relation expressed by an NR construction sounds weaker than a hypotactic realisation with a cue phrase. Therefore, if a relation is to be emphasised, NR constructions should not be used. However, a non-finite NR form usually makes the semantic implicature more explicit than the finite form (Quirk et al., 1985), which could be yet another alternative realisation. For example, comparing (5.7b) with (5.7c), the causal relation expressed in the latter is more explicit than that in the former. However in this case, it is usually not clear whether the non-restrictive clause is a postmodifier of the NP or an adverbial clause of the sentence.

(5.7) a. *Private Eye had been threatened with closure because it couldn't afford the libel payment.*

b. *Private Eye, which couldn't afford the libel payment, had been threatened with closure.*

c. *Private Eye, not being able to afford the libel payment, had been threatened with closure.*

5.4 Summary

This chapter investigates a specific type of NP modifier usage, the use of non-referring components to express semantic relations other than ELABORATION. This is a commonly used strategy by human authors, but has not been explored by an NLG system before. We focus on two semantic relations signalled by *because* and *then* respectively and describe an experiment to test the acceptability of NR constructions for expressing these relations. The experiment tests several relevant factors and enables us to accept or reject a number of hypotheses. It shows that *inferrability* does play a role in

realising the causal relation and when the conditions for *inferrability* etc. are satisfied, certain relations can be expressed through an NR construction as well as a normally used hypotactic construction.

The experiment follows the growing trend of applying empirical methods to NLG and the result is intended to be reliable enough to be used by NLG systems in generating descriptive text. In Chapter 7, we will briefly introduce the implementation of these results in an NLG system.

Chapter 6

Aggregation and Text Structuring

Apart from interacting with low level generation tasks like referring expression generation as discussed in Chapter 3, aggregation also interacts with high level processes such as text structuring in a complex way. We argue that how to resolve the complex interactions within and between tasks is more important to the generation of a coherent text than how to model each individual factor of tasks. This chapter first describes the interactions between aggregation and text structuring in detail. Then it tries to capture the interactions discussed through Chapters 3, 5 and 6 as preferences among features considered by different tasks. Heuristics for the preferences are derived from general linguistic and discourse theories. These heuristics will be used in the implementations to be described in Chapter 7 and will be evaluated by considering the quality of the generated texts.

6.1 The Effect of Aggregation on Discourse Coherence

In descriptive text generation, the task of text planning is to select the relevant information to be expressed in the text and organise it into a hierarchical structure which captures certain discourse preferences such as preferences for the use of rhetorical relations and discourse topic moves. The structure should feature coherent connections between all of its sub-structures.

The task of aggregation is to combine simple representations to form complex ones, which in the mean time leads to a shorter text as a whole. Aggregation can affect the ordering of text plans and the length of a paragraph or the whole text. So it is closely related to text planning tasks like maintaining discourse coherence, i.e., text structuring. Again, we do not consider content determination in this thesis.

In this chapter, we first briefly introduce two types of discourse coherence, local coherence and global coherence or entity-based coherence and relation-based coherence, in Section 6.1.1. Then we discuss the complex interactions between aggregation and maintaining discourse coherence. We argue that taking aggregation into account in text structuring is important to planning both levels of coherence. The discussions in Chapters 3, 5 and 6 motivate a set of heuristics for capturing the preferences among features considered by aggregation and text structuring (Section 6.2). Unlike previous chapters which rely on empirical methods to derive rules and heuristics, this chapter sometimes uses our intuitions which are tested indirectly via the implementations. Finally in Section 6.3, we argue for a generation architecture that can capture all these interactions in a principled way. Most examples and semantic relations used in this chapter are from the ILEX system.

In the rest of this section, we introduce representative accounts of discourse structures and how aggregation affects the construction of a coherent discourse structure.

6.1.1 Two Types of Coherence

In the theory of discourse structure developed by Grosz and Sidner (1986), a discourse structure has three components:

- *a linguistic structure*, where a discourse is divided into discourse segments,
- *an intentional structure*, which comprises the intentions behind the discourse segments and the relations connecting the segments, and
- *an attentional state*, which models the discourse participants' focus of attention.

Discourse segments are connected by either a *dominance* relation or a *satisfaction-precedence* relation. Each discourse segment exhibits two types of coherence: *local*

coherence among utterances inside the segment, and *global coherence* between this segment and other discourse segments.

While this discourse model is popular among researchers working on discourse interpretation, it is not precise enough to attract those interested in multisentential discourse generation. It lacks clear definitions for many basic concepts, e.g., segments and the relation between segments. In the generation community, theories based on domain-independent rhetorical relations, in particular, Rhetorical Structure Theory (Mann and Thompson, 1987b) (introduced in Section 1.1.1), are often adopted. According to RST, a natural text can be described as a hierarchical structure with a nucleus-satellite or multi-nuclear relation between every two sister spans of the text. So discourse coherence is achieved by connecting text spans with proper rhetorical relations. This is sometimes called *relation-based coherence.*

There has been an effort to synthesise the two accounts of discourse structure. Mann and Thompson point out that the two theories "are strongly related, partly because they share several important assumptions about the nature of the use of language and how to account for it". In general, RST produces a finer-grained account for discourse coherence than Grosz and Sidner's theory does. Moser and Moore (1996) argue that the two theories have considerable common ground, which lies in the correspondence between the notion of dominance and nuclearity. It is possible to map between Grosz and Sidner's linguistic structure and the RST tree structure. Therefore, RST relations refine the *dominance* and *satisfaction-precedence* relations between segments, and relation-based coherence and global coherence capture similar discourse properties.

However, the relation set proposed in RST is composed of heterogeneous relations. In addition to the distinction between subject-matter/informational relations and presentational/intentional relations mentioned in (Mann and Thompson, 1987b; Moore and Pollack, 1992), Knott et al. (in press) in particular discuss the incompatible properties of the OBJECT-ATTRIBUTE ELABORATION relation, which can hold between any two text spans about a common object, with respect to other RST relations. They propose a new discourse model which preserves the RST-based model except for the removal of OBJECT-ATTRIBUTE ELABORATION.

The new model also claims two types of coherence similar to global and local coherence, but it distinguishes itself in the way it interprets them. In this model, global coherence is featured as a sequence of focus spaces (discourse segments) about global foci, connected by a *resumption* relation, which exists between two segments where a later segment is about an entity in an earlier one. A resumption relation represents a global focus shift between focus spaces (named as *entity-chains*). Local coherence inside an entity-chain is however featured as both RST style trees capturing coherent connections between some propositions and a connection similar to OBJECT-ATTRIBUTE ELABORATION between other propositions.

This model is motivated by a study of descriptive text, in particular museum descriptions, which proves to be difficult for a purely RST-based interpretation. The development of the model is still in an initial stage and there is no clear intention from its authors to extend it to other domains. While the argument about what is an appropriate account for global and local coherence continues, we simply adopt the distinction between *entity-based coherence*, which exists between text spans in virtue of shared entities, and *relation-based coherence*, which exists between text spans connected by RST relations except for OBJECT-ATTRIBUTE ELABORATION, without claiming which accounts for the general or local coherent organisation of a discourse. We also take the point that the effect of OBJECT-ATTRIBUTE ELABORATION can be achieved through a careful control of focus moves, as discussed in (Knott et al., in press).

To generate a coherent text, a text planner must try to achieve both entity-based and relation-based coherence during planning. In the following, we will discuss how aggregation affects the planning of these two types of coherence. Note that when we talk about the ELABORATION relation below, we actually refer to OBJECT-ATTRIBUTE ELABORATION.

6.1.2 Embedding and Entity-based Coherence

In a structured text plan produced by text planning, entity-based coherence is normally maintained through the ordering of the selected facts. In terms of Centering Theory (Grosz et al., 1995), the ordering prefers certain types of center transition (e.g., center continuation) over others (e.g., center shifting). Entity-based coherence is maintained

by only including a proposition if it is related to the previous discourse through a preferred center move.

Embedding may affect text structuring by taking away facts from a sequence featuring preferred center movements for embedding, so the possibilities for the entities inside these facts to be potential foci of later discourse are reduced. As a result, the preferred center transitions in the original sequence could be cut off. For example, comparing the two descriptions of a necklace in (6.1), (6.1b) is less coherent than (6.1a) because of the sudden shift from the description of the necklace to that of the designer, which is a side effect of embedding.

(6.1) a. *This necklace is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. It was designed by Jessie King. King was Scottish. She once lived in London.*

b. *This necklace,* ***which was designed by Jessie King,*** *is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. King was Scottish. She once lived in London.*

We mentioned in Section 3.3 that the centers of sentences are normally realised as NPs. Since embedding adds non-referring components into an NP, it could affect the way a *Cb* is realised. As pointed out in (Grosz et al., 1995), different realisations (e.g., pronoun vs. definite description) are not equivalent with respect to their effect on coherence. Therefore, embedding could influence entity-based coherence by forcing a different realisation from that preferred by Centering Theory. There is an obvious need to balance the consideration for coherence and stylistic issues.

These examples show that although embedding operates on NPs, its effect is not limited to NPs, but spreads to a wider text span. In Section 3.5, we give Rule 3.2, which says that adding a non-referring part to a referring expression should not reduce the readability of the text. One restriction concerning readability is that the generated referring expression should not be too complex to read. The above discussion leads to a second restriction concerning Rule 3.2, that is, embedding should not reduce the

entity-based coherence of a discourse. This includes both what is described above and the case where the embedded part is too complex and therefore affects focus continuation. For example,

(6.2) *This jewel was designed by Jessie King,* ***who was a famous Scottish jeweller, but worked in London.*** *It is made of silver and enamel.*

In (6.2), the complex embedded component distracts from the center continuation and makes the use of the pronoun in the second sentence awkward.

To conform to Rule 3.2 and be able to generate flexible non-referring NP components at the same time, the effect of embedding on entity-based coherence must be considered in text planning to maintain the readability of the generated text.

6.1.3 Aggregation and Relation-based Coherence

In Section 2.2, we defined semantic parataxis, which concerns facts related by explicit multi-nuclear semantic relations like SEQUENCE and CONTRAST or by implicit connections like parallel common parts. If two facts have at least two identical parallel components, we say that a CONJUNCT or DISJUNCT relation exists between them, depending on how the two facts are related. These relations are multi-nuclear relations. In this way, semantic parataxis can be treated as a combining operation on text spans connected by such a relation, just like hypotaxis. We are not interested in purely textual parataxis, e.g., using coordinators like *and* to combine adjacent utterances connected by the JOINT relation.

Different types of aggregation need to be coordinated in the production of a coherent text. Complex embedded components like non-restrictive clauses may interrupt the semantic connection or syntactic similarity between a set of clauses. For example, suppose we have a set of facts such as in (6.3a) and (6.4a). If we do not consider the semantic connections while making embedding decisions, we could generate sentences like (6.3b) and (6.4b) respectively, which are not good compared with (6.3c) and (6.4c) (this judgement was agreed by another person, a native speaker).

(6.3) a. (u_1) *This necklace is made of gold.* (u_2) *It is also made of sapphire.* (u_3) *It is*

also made of enamel. (u_4) *Sapphire is a kind of precious stone of a transparent bright blue colour.* (u_5) *Enamel is often used to produce a very shiny surface.*

b. *This necklace is made of gold, sapphire,* ***a kind of precious stone of a transparent bright blue colour****, and enamel,* ***which is often used to produce a very shiny surface.***

c. *This necklace is made of gold, enamel and sapphire. Sapphire is a kind of precious stone of a transparent bright blue colour, and enamel is often used to produce a very shiny surface.*

(6.4) a. (u_1) *This jewel was made by Bjorn Weckstrom.* (u_2) *Weckstrom's jewels are normally sold through a company called Lapponia Jewellery.* (u_3) *However, this jewel was sold directly to the museum.* (u_4) *The company was established in 1933 by a jeweller called G. Pedersen.* (u_5) *Pedersen specialised in Art-deco jewels.*

b. *This jewel was made by Bjorn Weckstrom. Weckstrom's jewels are normally sold through a company called Lapponia Jewellery,* ***which was established in 1933 by a jeweller called G. Pedersen, who specialised in Art-deco jewels****. However, this jewel was sold directly to the museum.*

c. *This jewel was made by Bjorn Weckstrom. Weckstrom's jewels are normally sold through a company called Lapponia Jewellery. However, this jewel was sold directly to the museum. The company was established in 1933 by a jeweller called G. Pedersen,* ***who specialised in Art-deco jewels****.*

Figure 6.1 illustrates the RST style trees of some of the above examples, where the utterances in brackets are not individual clauses any more, but become embedded or coordinated NP components (see the ILEX User Manual (Knott and O'Donnell, 1998) for a description of the semantic relations used in the examples). The major differences between the two corresponding paragraphs are the embedded components in the middle of both the semantic parataxis in Example (6.3) and the expression of CONCESSION in Example (6.4), which we claim are the reasons for the less coherent paragraphs. Note that the figure only gives one possible tree for each example.

In contrast, adjectives would not have such negative effect in most cases, especially

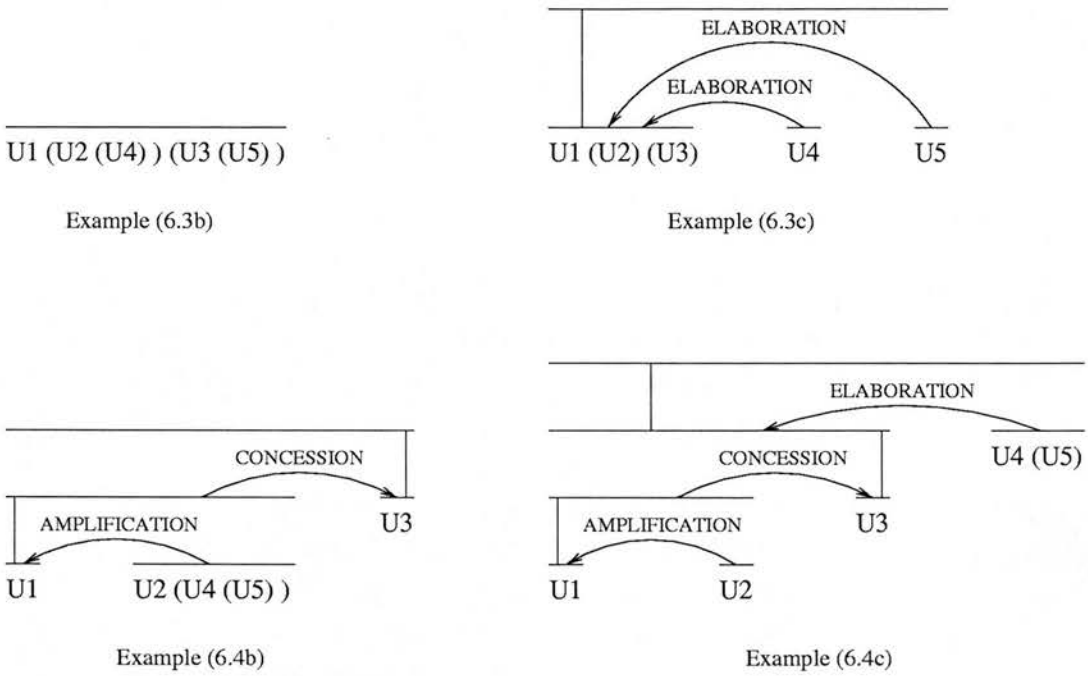


Figure 6.1: A comparison of the RST style trees for Examples (6.3b) and (6.3c), and for Examples (6.4b) and (6.4c)

when the paratactic parts have syntactically symmetrical modifications, such as Example (6.5).

(6.5) *This bracelet has a **slightly flared** band and a **swelling** midsection.*

The above problem is not only for embedding inside semantically related facts, but also for embedding of such facts. If one fact is to be embedded, so are all its semantically related facts if the relation is to be expressed in the text. That is, the possibilities of other types of aggregation should be considered for both the main fact and the fact to be embedded during embedding decision making. For example, when using embedding on the set of sentences in (6.6a), the realisation in (6.6b) rather than that in (6.6c) should be generated.

(6.6) a. *The necklace is in the Organic style. It is made of gold. It is also made of enamel.*

b. *The necklace, **which is made of gold and enamel**, is in the Organic style.*

c. *The necklace, **which is made of gold**, is in the Organic style. It is also made of enamel.*

In summary, if a hypotaxis or parataxis is to be performed for communication reasons, embedding should only be considered when it does not have a negative effect on these two types of aggregation so that the interpretation of the underline relation will not be interfered. That is, the embedded component should not lie between propositions that can be aggregated in such ways or consume facts required for such aggregation. This results in a third restriction concerning Rule 3.2, that is, embedding should not reduce the relation-based coherence of a discourse.

The complete Rule 3.2 is given below as Rule 6.1.

Rule 6.1 *The non-referring part should not reduce the readability of the text. This includes three aspects:*

The referring expression should not be too complex to read.
The non-referring part should not reduce the entity-based coherence of a discourse.
The non-referring part should not reduce the relation-based coherence of a discourse.

However, if the embedded material supports the semantics of the main proposition in some way, such as that described in Chapter 5, it probably will not interfere with relation-based coherence. Here is another example. Comparing the two embeddings in (6.7), (6.7a) is not as good as (6.7b) because the second non-restrictive clause provides evidence to increase the credibility of the main clause, whereas the first one only presents more information about a domain object.

(6.7) a. *If Mary ever has an opportunity to go to Paris,* ***which is the capital city of France****, she will stay there.*

b. *If Mary ever has an opportunity to go to Paris,* ***which has been her dream city for many years****, she will stay there.*

Apart from the aspects discussed in Chapter 5, it is unclear how such preferences can be used in current NLG systems.

In addition, performing parataxis inside a hypotaxis could convey wrong information.

For example, if we have a set of sentences like (6.8a), it would be wrong to say (6.8b) instead. The difference is illustrated in Figure 6.2, where (u_4) is aggregated into the AMPLIFICATION relation in (6.8b). This is a serious problem for a similarity based aggregation approach, which does not consider the hierarchical organisation of a discourse.

(6.8) a. (u_1) *The necklace is set with jewels* (u_2) *in that it features cabuchon stones.* (u_3) *Indeed, an Arts and Crafts style jewel usually uses cabuchon stones.* (u_4) *An Arts and Crafts style jewel usually uses oval stones.*

b. *The necklace is set with jewels in that it features cabuchon stones. Indeed, an Arts and Crafts style jewel usually uses cabuchon stones and oval stones.*

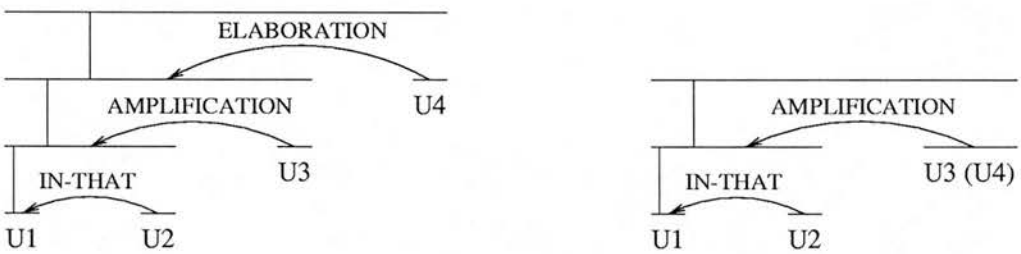


Figure 6.2: A comparison of the RST style trees for Examples (6.8a) and (6.8b)

We mentioned in Section 2.4 that in many domains it is potentially dangerous to allow parataxis to redo earlier decisions (e.g., fact ordering) because such operations might result in serious destruction to the satisfaction of some goals. In this case, semantic parataxis depends on the text planner to put the related facts next to each other during text planning in order to perform a combination. This makes it obvious that some paratactic possibilities must be considered in text structuring.

6.1.4 Aggregation and Paragraphing

Although there has been much research into the desirable extent of a paragraph from different perspectives, there is no general agreement as to what material should belong to a paragraph.

Zadrozny and Jensen (1991) classify the approaches to studying paragraphs into four groups:

- *Prescriptivist*: as in a standard English grammar book, where a paragraph is a group of sentences developing a topic;
- *Psycholinguist*: "a paragraph is a psychologically real unit of discourse, and, in fact a formal grammatical unit";
- *Textualist*: a paragraph is a hierarchical organisation of sentences;
- *Discourse-oriented*: the paragraph is the basic unit of semantic processing.

The prescriptivist's view of paragraph is most relevant to generation. According to this view, the information composing a paragraph must stick together and all be related to a topic (Strunk and White, 1979). The topic should either already exist in the background knowledge of a paragraph or it must be introduced in a sentence. However, since there is no clear definition of topic or subtopic, it is difficult to derive discourse structure from topic structure.

Research in hypertext and text display has produced hypotheses about how textual information should be displayed to users. Girill (1991) compares display of fine-grained portions of text (sentences), full texts and intermediate-sized units of an on-line documentation system, and finds that divisions at the fine-grained level are less efficient to manage and less effective in delivering useful answers than intermediate-sized units of text. But he does not make a commitment about exactly how large the desired text unit should be. The implication is that the proper unit is the one that groups together the information that performs some communicative function, which will range from one to several paragraphs.

Little research addresses the problem of how to decide appropriate paragraph boundaries in NLG. In most cases, paragraph structure is determined by high level schemata, which organise, integrate and predict material in the process of text construction (Dillon, 1981). It is not clear how paragraph structure can be decided in a text planner without schemata.

In general, paragraphing is a very complex issue. In our domain, the text structure is relatively flat and there is no obvious topic sentence. So the paragraphing issue might be simpler. If we adopt the prescriptivist's view of paragraph and assume that each

paragraph in descriptive text is about a domain object, then aggregation can affect paragraphing because it integrates information together and often reduces the number of sentences that could form a paragraph. As a result, many small paragraphs could be produced. If the topic changes constantly from paragraph to paragraph, the text would be incoherent, for example,

> *This jewel is a bracelet, which is 0.6 cm wide. It was made by a* ***famous young English*** *woman called Gerda Flockinger,* ***who lived in London.*** *It has a slightly flared band and a swelling midsection. This bracelet, which was made in London, is in the Organic style. It was made in 1965. It is made from pearl, aquamarines, turquoise, tourmalines and oxidised white metal. It draws on natural themes for inspiration in that it is a remarkably fluid piece. Indeed Organic style jewels usually draw on natural themes for inspiration.*
>
> *Flockinger was one of the best jewellers working in this medium.*
>
> *Organic style jewels are usually encrusted with gems and made up of asymmetrical shapes. They usually have a coarse texture.*

Although each paragraph above is coherent, the text as a whole is not because there are several small paragraphs. Some of them are formed because of the small number of facts concerning a topic, whereas others are caused by embedding consuming facts about a topic (as shown by the parts in boldface). These small paragraphs follow one another and are distracting. In addition, although the small paragraphs are related to the previous discourse through the *resumption* relation (Knott et al., in press), there are no explicit connection and transition markers.

To avoid this side effect, aggregation, in particular embedding, should be considered in text structuring when deciding paragraph boundaries, that is, removing possible small paragraphs by embedding them inside bigger ones. If however embedding would result in a small paragraph, say containing only one or two propositions, it should probably be avoided. This would then produce different planning sequences. For example, a reorganisation of the above example looks like:

> *This jewel is a bracelet, which is 0.6 cm wide. It has a slightly flared*

band and a swelling midsection. It was made in 1965 and was made in London. It is made from pearl, aquamarines, turquoise, tourmalines and oxidised white metal.

The bracelet was made by a famous young English woman called Gerda Flockinger, who lived in London. Flockinger was one of the best jewellers working in this medium.

The bracelet is in the Organic style. It draws on natural themes for inspiration in that it is a remarkably fluid piece. Indeed Organic style jewels usually draw on natural themes for inspiration. Organic style jewels are usually encrusted with gems and made up of asymmetrical shapes. They usually have a coarse texture.

This is more coherent than the previous version and the paragraphs seem to get a balanced size as a side effect. Note that only the second paragraph is a direct result of taking into account embedding in discourse structuring, but the point is that aggregation represents one factor that needs to be considered in text planning.

The discussion in this whole section shows that the effect of aggregation is not limited to the particular phrase or sentence where aggregation happens, but to the coherence of the text as a whole. Together with the discussion in Chapters 3 and 5, we argue that the complex interactions demand the features of aggregation to be evaluated together with other coherence features and aggregation to be planned as a part of text structuring. This requires better coordination between aggregation and other generation tasks as well as among different types of aggregation than is present in current NLG systems. In the following, we will put the bits and pieces discussed so far together and show how to use them as constraints on text production.

6.2 Capturing the Interactions as Preferences

It is obvious that a coherent text cannot be generated by just considering a single factor for coherence because many factors interact with each other in a complex way and together contribute to the goal. We claim that it is the relative preferences among features rather than the absolute magnitude of each individual one that play the crucial

role. Therefore, if we can capture these preferences in a generation system properly, we will be able to produce coherent text.

In this section, we first discuss the preferences among features related to text structuring, based on which those for aggregation can be introduced. In Chapter 7, we will describe how to implement these preferences in a pipeline and a non-pipeline generation architecture to build text structures that are coherent on both an entity and a relation basis.

6.2.1 Preferences among Coherence Features

Based on the current understanding of text coherence, it is impossible to give a complete list of preferences that guarantees a coherent text. So what we will present is a subset of preferences that we have observed in museum descriptions. To make them work in a different domain, substantial adaption or extension might be needed. We hope that the subset represents a starting point towards the extraction of a complete set.

For relation-based coherence

We have mentioned that a major semantic relation in descriptive text is OBJECT-ATTRIBUTE ELABORATION. Based on the discussion in Section 6.1.1, we do not consider it as an explicit relation, but rather assume the strategy of (Mellish et al., 1998a) which uses a JOINT relation to connect every two text spans that do not have a normal semantic relation or a CONJUNCT or DISJUNCT relation (defined in Section 6.1.3) in between. In the following, we use semantic relation to refer to relations other than JOINT, CONJUNCT and DISJUNCT.

A semantic relation is preferred to be used whenever possible because it usually conveys interesting information about domain objects and leads to a coherent text span. How to choose among several possible relations is the concern of the text planner and should usually be computed considering domain specific goals. However, a semantic relation can only be used if all presuppositions of that relation are satisfied, that is, the knowledge assumed to be shared by the hearer is introduced in the previous discourse.

If the presuppositions are not met, a semantic relation should not be used because it would probably confuse the reader. For example, the AMPLIFICATION relation between the propositions in (6.9a) can only be introduced if (6.9b) has been mentioned in the previous discourse. Otherwise, the reader might have difficulty to infer the connection between the two propositions.

(6.9) a. *This necklace has silver links encrusted asymmetrically with pearls and diamonds. Indeed, Organic style jewels are usually encrusted with jewels.*

b. *This necklace is in the Organic style.*

If a CONJUNCT or DISJUNCT relation shares a fact with a semantic relation, it should be suppressed because a semantic relation is thought to convey more interesting information about domain objects. For the set of utterances in Example (6.10a), apart from other relations, there is an AMPLIFICATION between (u_3) and (u_4) and a CONJUNCT between (u_4) and (u_5) to choose from, both of which make use of (u_4). Compared with (6.10a), (6.10b) is less preferred because it misses the AMPLIFICATION relation and the transition from the description of *the necklace* to that of *an Arts and Crafts style jewel* is not so smooth, whereas (6.10a) expresses AMPLIFICATION explicitly but misses out the CONJUNCT. The difference is shown by a possible tree of each example in Figure 6.3.

(6.10) a. (u_1) *The necklace is in the Arts and Crafts style.* (u_2) *It is set with jewels* (u_3) *in that it features cabuchon stones.* (u_4) *Indeed, an Arts and Crafts style jewel usually uses cabuchon stones.* (u_5) *It usually uses oval stones.*

b. *The necklace is in the Arts and Crafts style. It is set with jewels in that it features cabuchon stones. An Arts and Crafts style jewel usually uses cabuchon stones and oval stones.*

Although JOINT is not preferred when other relations are present, it is better than relations with missing presuppositions or expressing a CONJUNCT inside a semantic relation (as introduced in Section 6.1.3, in particular, Example (6.8)). Therefore, we have the following heuristics, where "A$\prec$B" means that A is preferred over B.

Heuristic 6.1 *Preferences among features for relation-based coherence:*

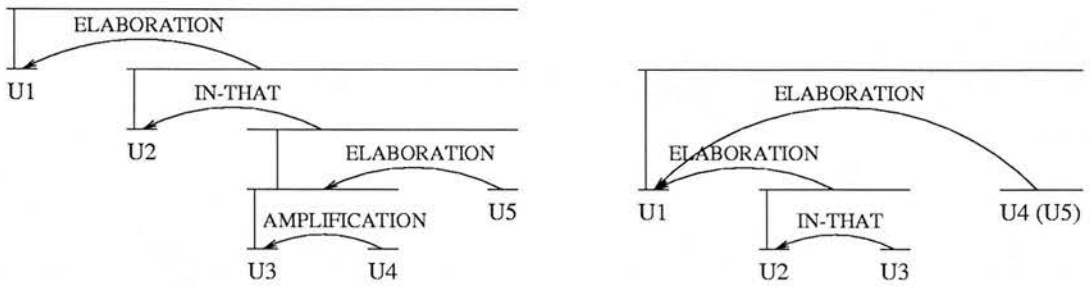


Figure 6.3: A comparison of the RST style trees for Examples (6.10a) and (6.10b)

a semantic relation $\prec$ CONJUNCT/DISJUNCT $\prec$ JOINT $\prec$ *presuppositions of a relation not met*

JOINT $\prec$ CONJUNCT/DISJUNCT *inside a semantic relation unless the relation holds for all conjoined facts*

For entity-based coherence

One way to achieve entity-based coherence is to control center transitions among utterances. In Centering Theory, Rule 2 specifies preferences among center movements in a locally coherent discourse segment: sequences of *continuation* are preferred over sequences of *retaining*, which are then preferred over sequences of *shifting*.

Brennan et al. (1987) also describe typical movements of a discourse topic in terms of center transitions between pairs of utterances. They are given in Table 6.1, where *Cp* is the preferred center, the highest ranked element among the forward-looking centers. They argue that if a speaker has a number of propositions to express, the simplest way to do so coherently is to express all the propositions about a given entity (*continuing*), before introducing a related entity (*retaining*), and then shifting to this new entity. The order of coherence among the transitions is *continuing* $\prec$ *retaining* $\prec$ *smooth shifting* $\prec$ *abrupt shifting*, where those in the front are more coherent.

	Cb(Un)=Cb(Un+1)	Cb(Un)≠Cb(Un+1)
Cb(Un+1)=Cp(Un+1)	continuing	smooth shifting
Cb(Un+1)≠Cp(Un+1)	retaining	abrupt shifting

Table 6.1: Typical focus moves defined in (Brennan *et al.* 87)

Instead of claiming that these are the best models, we use them simply as an example of linguistic models being used for evaluating factors for text structuring.

A type of center transition that appears frequently in descriptive text is that the description starts with an object, but shifts to associated objects or perspectives of that object. For example, when we describe a car, we start with the car itself and then move on to talk about its engine. This is a type of abrupt shifting according to (Brennan et al., 1987), but it is appropriate as long as the objects are highly associated to the original object (Schank, 1977). This phenomenon is handled in a number of systems. For instance, Grosz (1977) includes the subparts of an object into a focus space when the object itself is to be included. The subparts are defined as the implicit foci, which could be used to constrain what can be said next. McCoy and Cheng (1991) use discourse focus trees to capture the possible change of focus among objects, where object perspectives (including associated objects and attributes) are obvious candidates.

We call this center movement an *associate shifting*, where the center moves from a trigger entity to a closely associated entity, which is usually expressed by a bridging description (introduced in Section 4.4.1). There are two types of associate shifting: the trigger is in the previous utterance or two entities in two adjacent utterances have the same trigger. Our informal observation from museum descriptions shows that an associate shifting is preferred by human writers to all other types of center movements except for a continuation.

Heuristic 6.2 below summarises the preferences among the different types of center transition described above. We acknowledge that this is a strict heuristic and that human written texts are sometimes more flexible.

Heuristic 6.2 *Preferences among center transitions:*

$$\textit{Continuation} \prec \textit{Associate shifting} \prec \textit{Retaining} \prec \textit{Smooth shifting} \prec \textit{Abrupt shifting}$$

As a result, semantic relations connecting propositions with smooth center transitions would be preferred over those without. In Example (6.11), (6.11a) and (6.11b) express

two different semantic relations which consume a common proposition. If the two relations are equally interesting and only one can be included, (6.11b) would be preferred because it features a center continuation in addition to a semantic relation.

(6.11) a. *This necklace draws on natural themes for inspiration. Indeed, an Organic style jewel usually draws on natural themes for inspiration.*

b. *This necklace draws on natural themes for inspiration, for example, it uses natural pearls.*

For relation-based and entity-based coherence

Two propositions can be connected in different ways, e.g., through a semantic relation or a smooth center transition only. Since a semantic relation is preferred to be used whenever possible, we have the following heuristic:

Heuristic 6.3 *Preferences among semantic relations and center transitions:*

a semantic relation + Abrupt shifting $\prec$ JOINT *+ Continuation*

This means that we assume semantic relations dominate center transitions in our domain.

6.2.2 Preferences among Embedding Features

In Section 6.1, we describe the effect of embedding on entity-based and relation-based coherence. The complex interaction demands the features of embedding to be evaluated together with other coherence features so that embedding can be planned as a part of text structuring. This demand can be satisfied by assuming that embedding is an alternative to using a JOINT relation between two propositions with a common entity. We need to find out when embedding is a better alternative.

We distinguish between a *good, normal* and *bad* embedding based on the features it bears. We do not claim that the set of features is complete. In a different context, more criteria might have to be considered. So *good, normal* and *bad* are judged within the features we consider in this thesis.

A good embedding is one satisfying all the following conditions, which have been discussed in Chapters 3 and 5:

1. The referring part is an indefinite phrase, a demonstrative phrase, a proper name, a bridging description where the cardinality of the association relation is one, or a definite description where the referent might have to be the most salient object in the context.

2. The embedded property can be realised as an adjective (not for proper names), a prepositional phrase, an appositive component, or a non-restrictive clause with a causal relation (of strong inferrability) or a temporal relation between the main clause and the non-restrictive clause.

3. The embedded property is not in a semantic relation or in a CONJUNCT/DISJUNCT relation with another fact. And in the resulting text, the embedded part does not lie between text spans connected by one of these relations.

4. There is an available syntactic slot to hold the embedded part.

A good embedding is highly preferred and should be performed whenever possible. A normal embedding is one satisfying conditions 1, 3 and 4 and the embedded part is a relative clause which provides additional information about the referent denoted by the referring part. Since there is still a lack of effective way of testing the relevance between the main clause and the relative clause, we do not treat it as very bad if there is no relevance. Bad embeddings are all those left, for example, if there is no available syntactic slot for the embedded part.

As argued in Section 6.1.3, semantic parataxis has a higher priority than embedding. Therefore, a good embedding should be less preferred than using a CONJUNCT relation.

To decide the interaction between an embedding and center transitions, we include Example (6.1) here again as (6.12). The only difference between (6.12a) and (6.12b) is the position of the sentence "*This necklace was designed by Jessie King*".

(6.12) a. *This necklace is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance,*

this necklace has floral motifs. It was designed by Jessie King. King was Scottish. She once lived in London.

b. *This necklace,* ***which was designed by Jessie King****, is in the Arts and Crafts style. Arts and Crafts style jewels usually have an elaborate design. They tend to have floral motifs. For instance, this necklace has floral motifs. King was Scottish. She once lived in London.*

The difference can be represented in terms of features of entity-based coherence and embedding as follows:

the last three sentences in (6.12a): JOINT + *Continuation* + JOINT + *Smooth shifting* + JOINT + *Continuation*

the last two sentences plus embedding in (6.12b): JOINT + *Abrupt shifting* + *Normal embedding* + JOINT + *Continuation*

(6.12a) is preferred over (6.12b) because the center moves more smoothly in (6.12a). The heuristics derived from the above discussions are summarised below:

Heuristic 6.4 *Preferences among features for embedding and center transition:*

Good embedding $\prec$ *Normal embedding* $\prec$ JOINT $\prec$ *Bad embedding*

Continuation + *Smooth shifting* + JOINT $\prec$ *Abrupt shifting* + *Normal embedding*

Good embedding $\prec$ *Continuation* + JOINT

CONJUNCT $\prec$ *Good embedding*

6.2.3 Summary of Preferences

The preferences we have so far are summarised below. The '+' symbol can be interpreted in different ways, depending on how the features are modelled in NLG systems. In a traditional pipeline architecture, it means the coexistence of two features. In a system using numbers for planning, it can have the same meaning as the arithmetic symbol.

1. preferences among discourse relations:

   (a) *a semantic relation* $\prec$ CONJUNCT = DISJUNCT $\prec$ JOINT $\prec$ *presuppositions of a relation not met*

   (b) JOINT $\prec$ CONJUNCT/DISJUNCT *inside a semantic relation unless the relation holds for all conjoined facts*

   (c) *a semantic relation + Abrupt shifting* $\prec$ JOINT + *Continuation*

2. preferences among center transitions: *Continuation* $\prec$ *Associate shifting* $\prec$ *Retaining* $\prec$ *Smooth shifting* $\prec$ *Abrupt shifting.*

3. preferences among embeddings and center transitions:

   (a) *Good embedding* $\prec$ *Normal embedding* $\prec$ JOINT $\prec$ *Bad embedding*

   (b) *Continuation + Smooth shifting* + JOINT $\prec$ *Abrupt shifting + Normal embedding*

   (c) *Good embedding* $\prec$ *continuation* + JOINT

   (d) CONJUNCT $\prec$ *Good embedding*

In Chapters 7 and 8, two implementations using these preferences and their evaluation will be described.

6.3 Further Discussion

In Chapter 3, we argue that embedding is closely related to low level generation tasks such as referring expression generation and they interact in a complex way. This interaction demands that both tasks be considered at the same time rather than sequentially in order to generate referring expressions capable of serving multiple communicative goals. In our heuristics, although we define good embedding in terms of syntactic slots, this does not force embedding to be done after referring expression generation.

In Chapter 5, we describe `int` type modifiers, especially those supporting the semantics of the main propositions. The selection of such modifiers should be a concern of text planning. Since embedding concerns the realisation of all non-referring NP components,

it can suggest to the text planner as to whether a property can be realised through NP subordination, under the constraints from the NP type and the `unique` modifiers that are already there. Therefore, the generation of `int` modifiers needs the coordination between text planning and embedding. These are also considered in the definition of good embedding.

In this chapter, we describe the effect of aggregation on both entity-based and relation-based coherence. We show that embedding may affect center transitions and the construction of paragraphs. One topic we mentioned repeatedly during these discussions is that there are complex interactions between aggregation and other generation tasks, so aggregation should be taken into account during text structuring in order to produce a more coherent text.

As introduced in Chapter 2, aggregation is often performed on various forms of structured text plans from text planning, e.g., (Hovy, 1993). This is because in automatic NLG, various versions of the pipeline architecture specified by Reiter and Dale (Reiter, 1994; Reiter and Dale, 1997) are normally adopted. In this type of architecture, the output of text planning is normally a tree structure resembling an RST tree. Aggregation operates on such trees to combine subtrees with identical components. Pipeline architectures successfully modularise the individual generation problem, but fail to capture the complex interactions between different modules. The interactions we have described require an architecture that provides better coordination between aggregation and other generation tasks as well as among different types of aggregation than they are in current NLG systems.

This motivates the implementations to be described in the next chapter, where we will introduce the implementation of aggregation in both a pipeline and a parallel architecture. We wish to show that the interactions can be captured more naturally in the parallel architecture.

6.4 Summary

This chapter discusses the complex interactions between aggregation and high level generation tasks such as text structuring as well as the interactions between different

aggregation subtasks. Together with what is discussed in Chapters 3 and 5, they motivate a set of preferences among coherence features, which we claim capture some important properties that characterise a good descriptive text. They are also the motivations of a non-pipeline generation architecture which can model the interactions better. The embedding rules and heuristics from Chapters 3, 4, 5 and 6 are summarised in Appendix A.1.

Chapter 7

Implementing Aggregation in Two NLG Systems

The central issues of this thesis are three fold: revealing the interactions between aggregation and other generation tasks, extracting heuristics from these interactions for controlling the production of a coherent text and exploring the ways to model these heuristics in text generation systems. In the previous chapters, we have introduced the aggregation phenomena observed from our corpus and the preferences among features of aggregation and other generation tasks such as referring expression generation and text structuring. This chapter describes the implementation of these preferences in two text generation systems: ILEX-TS and GA-plan. We start with an introduction to text planning architectures, which features the major difference between the two systems, and then describe each specific implementation in more detail. To facilitate expressing the necessary syntactic restrictions, a revised version of Meteer's Text Structure is introduced as the intermediate representation between text planning and realisation. Comparisons of the generated texts are left for Chapter 8.

7.1 Text Planning: a Brief Introduction

In Chapters 3 and 6, we discuss the interactions between aggregation and other generation tasks and give heuristics for preferences among relevant features. We believe

that these heuristics capture some important coherence properties of at least a type of text, descriptive text, and if they can be modelled properly in a generation system, they will lead to coherent text. In this chapter, we describe the implementation of the heuristics in two generation systems. This enables an evaluation of the heuristics using the implemented systems and their output. The reason for two implementation is to compare different ways of modelling the interactions and find evidence concerning the performance of a pipeline architecture versus a parallel architecture in this respect.

Again we focus on such generation tasks as aggregation, document structuring and referring expression generation, and only briefly mention other tasks including content determination and realisation. We are also concerned with choosing the right architecture that could allow us to take into account the interactions, not just between modules but also among subtasks inside a module.

The architecture of text planning has a great effect on aggregation possibilities. Generally, an architecture that allows more than one text structure with the same degree of coherence to be built and more interactions between generation considerations to be incorporated is preferred over one that only produces limited types of text plan. To experiment with the complex interactions we have described, we need a generation architecture that is suitable for producing descriptive text as well as allowing a relatively large space for aggregation. In this section, we introduce three representative text planning strategies and discuss their appropriateness to studying aggregation and the task of descriptive text generation.

7.1.1 Top-down and Bottom-up Planning

Much research in text planning is based on domain-independent top-down planning strategies as proposed by Hovy and Moore.

(Hovy, 1988; Hovy, 1989) propose an RST-based text planning strategy using rhetorical relations as plan operators. Each operator has suggested *growth points*, which can be satisfied by selecting propositions from the knowledge base or become further planning goals. The planning process continues until no more goals to be satisfied or input units to be related.

Top-down planning usually results in an intermediate structure like an RST tree representing a coherent text plan. Aggregation works on this structure to combine adjacent sub-structures that are connected by certain rhetorical relations (such as LIST and ELABORATION) and have identical parts, e.g., (Dalianis, 1996; Huang and Fiedler, 1996). Aggregation considerations are not given high priority in text structuring and the space of aggregation is very limited, i.e., it can only happen on adjacent subtrees.

In the widely cited planning strategy for dialogue generation as proposed in (Moore and Paris, 1994), information about the intended effect of parts of the text on the hearer is included in the discourse model in addition to considering the rhetorical relations between parts of a text. This makes the dialogue system capable of reasoning about its previous utterances and interpreting follow-up questions in the context of a conversation. However, this strategy suits a domain with various obvious communicative intentions, not a domain like descriptive text whose major intention is simply to provide the hearer with objective and interesting information about the objects being described.

Kittredge et al. (1991) casts doubt on the suitability of the RST-based domain-independent planning for some kinds of text, like those in reporting domains. Among the problems he addresses, two are closely related to descriptive text generation:

- Descriptive text uses a JOINT schema frequently, which is a rather domain-specific relation and does not have general constraints on its nuclei. It is of no use to RST-based top-down planning;
- Growth point theory requires a relatively large number of relations presented in the domain text to narrow down the possible choices in the knowledge base, whereas descriptive text uses only a small subset of the subject-matter relations (in the sense of (Mann and Thompson, 1987b)).

(Marcu, 1997a; Marcu, 1997b) also discuss the inability of top-down planning to fulfill tasks like "tell everything that is in this knowledge base or everything that is in this chosen subset". Marcu formulates the task as constructing a text plan whose leaves subsume all the information given in a knowledge base, which consists of a set of

semantic units and rhetorical relations between pairs of units. He proposes a bottom-up, data-driven planning method using constraint satisfaction techniques, which tries to achieve global coherence through satisfying the local constraints on ordering and clustering of the nuclei and satellites of rhetorical relations. This approach is able to construct hierarchical text plans which satisfy multiple high-level communicative goals out of all the information in a knowledge pool.

A good example of aggregation in bottom-up style generation is presented in the Fragment-and-Compose paradigm of (Mann and Moore, 1981) (described in Section 2.3). In that architecture, aggregation is crucial to the formation of sentence-sized trunks from smaller fragments. It uses a set of aggregation rules resembling the clause-combining rules of English and a revised Hill Climber algorithm to search for the best application of the rules. In a way, the similarity-based aggregation strategies introduced in Section 2.4 can be seen as specific to bottom-up generation.

7.1.2 Opportunistic Planning

Based on Marcu's proposal, Mellish et al. (1998b) present an opportunistic planning paradigm, which aims at a text genre without an explicit overriding goal. The text planner postulates a number of interconnected communicative goals at each step of text planning, which mainly include expressing interesting information about the focussed entity and related entities within the domain. The planner tries to select facts to satisfy as many as possible of these interrelated goals and maintain a coherent connection between the current selection and the previous context in the mean time.

Both bottom-up and opportunistic planning suit a text genre that does not have a central overriding communicative goal which could be decomposed in a structured way into subgoals. The central goal of an object descriptive text is to provide interesting information about the target object. There are generally only a small number of relations, mainly OBJECT-ATTRIBUTE ELABORATION and JOINT, and sometimes COMPARISON and CONCESSION, etc. Most material is related by the elaboration relation. Therefore, a domain-dependent bottom-up or opportunistic planning strategy suits descriptive text generation better than a domain-independent top-down planning strategy.

Within this general framework, various implementation can be used. The planner can simply list all the materials about a domain object according to their importance to specific domain purposes or conventions and at the same time maintain smooth topic moves, or it can have complex heuristics as to which rhetorical relations are preferred and where to place them. In any of these cases, aggregation can be important to text planning because it changes the order in which information is expressed and affects topic moves. Therefore, an NLG system using a bottom-up or opportunistic planning strategy is suitable for studying aggregation in descriptive text generation.

In this chapter, we describe how the restrictions on the non-referring part (Rules 3.1 and 3.2) and the interactions between generation tasks (Heuristics 6.1 to 6.4) can be implemented in two generation systems: ILEX-TS and GA-plan. ILEX-TS is based on the ILEX system, which generates museum descriptions and uses a strategy combining bottom-up and opportunistic planning. However, we have to change the text planner slightly to give aggregation more attention. This implementation is described in detail in Section 7.3. GA-plan is an experimental generation system using a genetic algorithm for text structuring. It is chosen because it offers an excellent unconventional framework for modelling the interactions between generation tasks. More details are given in Section 7.4.

In the next section, we introduce a revised version of Meteer's Text Structure to provide abstract syntactic constraints on document structuring and referring expression construction. This gives an effective representation for taking into account realisation restrictions in content structuring and for controlling the complexity of the generated NPs as mentioned in Rule 3.2. The Text Structure is used in ILEX-TS.

7.2 Meteer's Text Structure

In a pipeline generation architecture, text planning is usually a self-contained procedure independent of linguistic realisation, whereas in fact, the choice of realisation forms affects the amount of information selected from the knowledge base and the relative salience of it. So there is a gap between the two processes in the pipeline architecture, which has been discussed by many researchers. Various methods have been proposed

to bridge the gap and Meteer's Text Structure (Meteer, 1991; Meteer, 1992) is a well-known example. The central idea is to propose an intermediate level of representation between text planning and realisation so that the planning process takes into account linguistic restrictions. Meteer designed a representation called Text Structure (TS), which is "an abstract linguistic level that reflects germane linguistic constraints while abstracting away from syntactic detail" (Meteer, 1991). In this section, we describe this representation as a way of providing linguistic constraints on embedding since we assume that reasoning about the syntactic requirement on embedding can be done using a revised version of Meteer's Text Structure.

7.2.1 Overview

Meteer's Text Structure is a tree structure with the nodes representing the constituents of a text and the utterances in it. It provides a unified representation for structures both above and within a sentence, so that document structuring and abstract sentence planning can be done at the same time. To abstract away the concrete lexical and grammatical detail, a node contains three major pieces of information:

- *Constituency*: the content of the utterance;
- *Structural relations among constituents*: the structural relations with respect to its parent and children;
- *Semantic category the constituent expresses*: the lexical head and the semantic category subsuming the constituent, which is used to constrain the expansion of the tree to include only expressible subtrees.

For each domain application object, a realisation class composed of resource trees is defined to build the possible Text Structures for that object. As the planning process progresses, the compatible resource trees of the domain objects are selected and inserted into the Text Structure, under the restrictions imposed by the structure that is already there. The Text Structure is traversed top-down and left-to-right twice: first to expand the tree structure and then to read out its elements. Below the Text

Structure is the linguistic specification which is used to map the utterances represented as Text Structure to linguistic surface structure.

(Panaget, 1994a; Panaget, 1994b; Panaget, 1997) improve Meteer's Text Structure by separating the semantic constraints into ideational and textual semantic constraints, using an *upper model* and a *hierarchy of textual semantic categories* respectively. The upper model is similar to the Generalized Upper-Model described in Section 3.2.2. The hierarchy of textual semantic categories is a domain independent inheritance network which organises concepts according to their textual realisation, e.g., *Time-anchored-clause, Intransitive-nucleus, Clause-modifier* (we simply use Panaget's term here although we think these are more like abstract syntactic categories). A fragment of the hierarchy is shown in Figure 7.1.

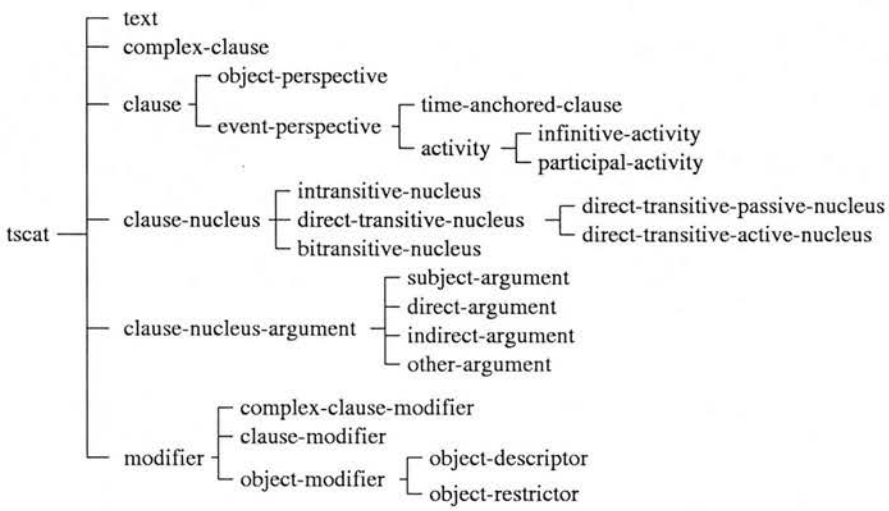


Figure 7.1: A fragment of the hierarchy of textual semantic categories

For each basic ideational category in the upper model, Abstract Linguistic Resources (ALR) encoding lexical and grammatical resources are defined to abstract away the linguistic detail. An ALR is a tuple: {*ideational-expr, textual-prop, surface-prop, constituency, constraints, style*}. The first three slots identify the ideational category in the upper model, the textual category in the textual semantic hierarchy and the surface properties of this resource respectively; *Constituency* defines the types of the sub-constituents of this ALR; *Constraints* are conditions that must be satisfied to select this ALR; *Style* is the stylistic features of this constituent. The idea is to represent

a linguistic resource as a structure containing information on its 'meaning', general textual property and specific textual property so that it can be used to restrict text planning. ALRs are used for building the Text Structure.

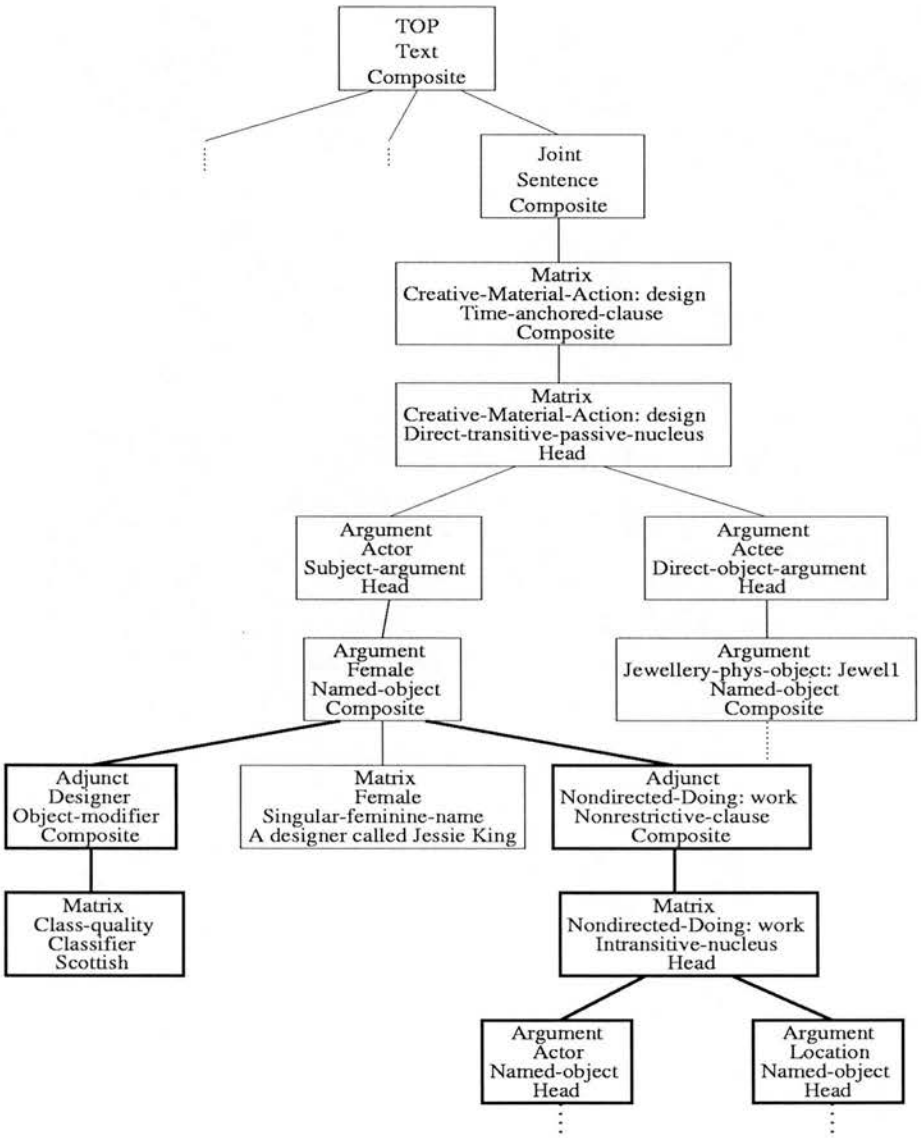


Figure 7.2: A fragment of the Text Structure for Sentence (7.1)

We adopt the improved Text Structure, i.e., Meteer's Text Structure using Panaget's two types of semantic categories, as our intermediate representation between text planning and realisation. Figure 7.2 is an example, which shows how a sentence and the embedded NP components are represented in the Text Structure (the constituency of the nodes are not shown). The structure without the highlighted parts is built by the

text planner without making embedding decisions. The highlighted parts are added by our embedding process. From this structure, Sentence (7.1) is generated.

(7.1) *This jewel was designed by a* ***Scottish*** *designer called Jessie King,* ***who worked in London.***

7.2.2 Why Use the Text Structure?

There are three reasons for the improved Text Structure to be an excellent intermediate representation for text planning and aggregation.

Firstly, many systems perform aggregation on some rhetorical hierarchical structure of text produced by top-down planning, where repetition is checked and substructures combined according to aggregation rules. We have mentioned in Section 7.1.2 that in object descriptions, most facts simply provide information for a few discourse entities and there are few different rhetorical relations between facts, usually ELABORATION, which connects facts about the same discourse entity. In this case, an ELABORATION can have a large number of possible expansions, which makes it hard for a domain independent top-down planner based on growth points (Hovy, 1988; Hovy, 1989) to determine which specific information to select from the knowledge base. A text planner based on goal decomposition, e.g., (Moore and Paris, 1994), may also be inadequate because there is no obvious goal structure in this genre.

A better way is to incrementally add and aggregate information in the knowledge base into an intermediate structure in a way that maximises conciseness and cohesion, that is, generating representations as concise as possible during document structuring rather than only checking for redundancy later on. "The Text Structure allows the generation process overall to be incremental, since it ensures that the text plan being composed will always be expressible in the language." (Meteer, 1991). So it is a suitable representation for aggregation in descriptive text generation.

Secondly, aggregation has a close relationship with the linguistic representation of sentences. As Robin (1993) states, "factors like conciseness and readability directly depend on surface form and monitoring them cannot be done by reasoning only at higher layers." A large part of what aggregation does is to change the inner structure

of clauses and produce a more complex clause structure including various nominal and verbal groups. In an RST tree, the smallest unit is a clause, a leaf of the tree. Text planning based on such a tree is not adequate for aggregation because the planning process stops at the clause level. In addition, although there are various sentence generation systems, it is doubtful whether they can realise sophisticated aggregation without syntactic representations. So what aggregation needs is a structured representation sensitive to linguistic structure, which is just what Text Structure is supposed to provide.

Finally, aggregation deals with combinations both within and between clauses. So it needs a representation which facilitates the same description below and above clauses, and Text Structure presents such a uniform representation for all the components of a text.

Lexicalised grammars such as Combinatory Categorial Grammar (CCG) (Steedman, 1996), Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Lexicalized Tree-Adjoining Grammar (LTAG) (Joshi and Schabes, 1992) can also provide the syntactic restrictions necessary for aggregation. However, there are two reasons that make such an option not suitable. Firstly, although there is no doubt that aggregation decisions cannot be made irrespective of syntactic constraints, it is not desirable to get into the full detail of syntactic realisation during aggregation and document structuring. What we need is a level of syntactic representation which is abstract in the sense that it does not directly encode such syntactic features as surface constituency and word order, etc. This is why an abstract syntactic representation is needed in our implementation and in NLG in general. The lexicalised grammars are too concrete for our purpose. Secondly, as mentioned above, aggregation needs a unified representation for both below and above the clause level, whereas these grammars are mostly focused on facilitating the syntactic structure within a clause.

Robin (1993) argues that "Meteer's Text-Structure remains too abstract. Although grammatical constituency is already decided at that level, many grammatical features and open-class lexical items with different stylistic impacts are not yet specified." Such demands on syntactic and lexical details may be due to the complexity of sports reports, whereas in our domain there is less emphasis on the quantity of information

inside sentences, but more on the coherence of the description as a whole, although complex NPs appear frequently. The Text Structure offers an integrated and unified representation for document structuring and aggregation and is sufficient for our domain.

In summary, the improved Text Structure satisfies the representation requirements of descriptive document structuring and aggregation. It guarantees that the structured text from the text planner can always be transferred to grammatically correct surface text.

7.3 Aggregation in ILEX-TS

7.3.1 An Overview of ILEX

We have mentioned that ILEX generates hypertext descriptions in the museum domain. It is capable of adapting its output to an initial communicative goal, the length of the output text, the profile of the user and discourse history.

In ILEX, pieces of domain knowledge that may be worth expressing in a text are represented as nodes and links in a graph called the *Content Potential*, which is compiled from a knowledge base of facts and rules and a user model. There are three kinds of nodes in the graph: entity nodes (each corresponding to an individual or generic domain object), fact nodes (each representing a relation between two entities, and expressible as a single sentence in language) and relation nodes (each representing a coherence relation between two facts, and expressible as a complex sentence in language). A fact is represented as *Predicate(Arg1,Arg2)*, where Arg1 and Arg2 are two entities.

Figure 7.3 from (Mellish et al., 1998b) illustrates what the Content Potential looks like. In addition to the three types of nodes we have just described, the links in the figure indicate the feature value connections between different types of nodes, e.g., the Arg1 role of a fact node connects this node with an entity node, and the Nuc (nucleus of an RST relation) role of a relation node links this node to a fact node.

ILEX uses a single knowledge representation formalism for both domain and textual

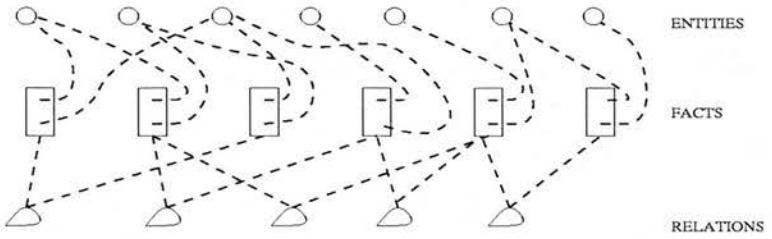


Figure 7.3: The Content Potential of ILEX

objects. Its knowledge representation system is built on top of WAG (Workbench for Analysis and Generation (O'Donnell, 1994)), which provides tools for representing and processing linguistic information represented in the Systemic formalism. WAG uses a set of *units* and *relations* between units to represent Systemic structures. A unit definition mainly contains the following fields:

`unit-id`	a unique identifier for the unit
`features`	a set of features for the unit
`roles`	the set of relations for which the unit is the head, along with the unit-id of the dependent of the relation (the filler)
`backpointers`	the set of relations for which the unit is the dependent, along with the unit-id of the head of each relation
`ordering`	orderings between this unit and other units
...	...

The nodes of the Content Potential are represented using the unit structure. In the next section, we will use it also to represent the nodes of Text Structure.

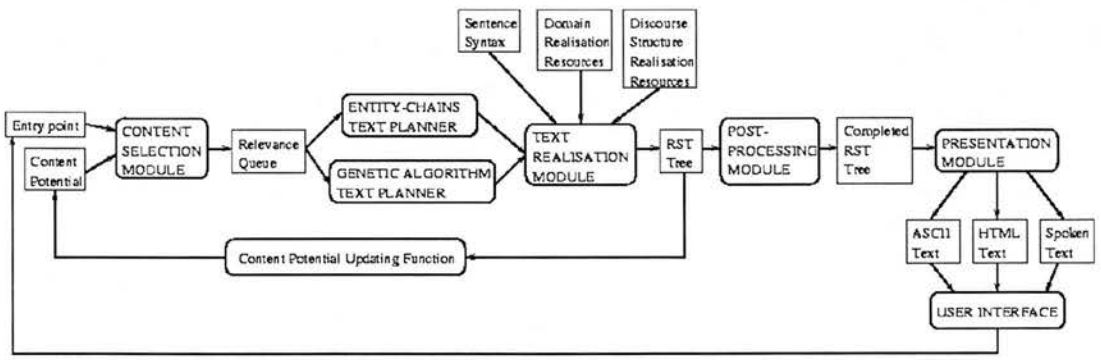


Figure 7.4: The ILEX architecture

Figure 7.4 from the ILEX user manual (Knott and O'Donnell, 1998) illustrates the ILEX generation architecture. The text planning task is fulfilled in two steps:

a content selection procedure, where a set of fact nodes with high relevance is selected from the Content Potential (following a search algorithm). This step follows the opportunistic paradigm described in Section 7.1.2, and is carried out in the Content Selection Module in Figure 7.4. The relevance is calculated from the importance and interestingness of facts as specified by domain experts and the weights of the relations between facts. The result is a relevance queue, a list of fact nodes ranked in order of decreasing relevance.

a content structuring procedure, where selected facts are reorganised to form entity-chains (based on the discourse theory of (Knott et al., in press)), which represent a coherent text arrangement. This step is carried out by the Entity-chains Text planner.

The structure of entity-chains is illustrated in Figure 7.5 from (Knott and O'Donnell, 1998). Entity-chains are basically sequences of text spans about a given entity (the focus). Text spans can be simple facts or RST trees, and they feature center continuations inside a sequence. Each entity-chain forms a paragraph. Between sequences of entity-chains, a resumption relation is used to form a coherent connection. A resumption relation applies between two chains C1 and C2 iff the focus of C2 is an entity that is mentioned at some point in one of the spans in C1. This is a type of smooth topic shift in a global sense. The ILEX text planner tries to build a structure consisting of a sequence of entity-chains where there is always a resumption relation between a chain and some chain to its left, except for the first one. This is a data-driven strategy combining opportunistic and bottom-up planning.

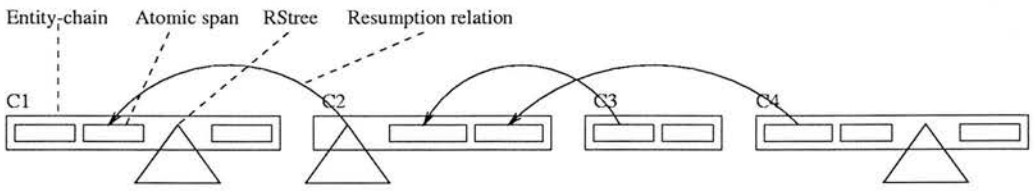


Figure 7.5: An illustrative structure of entity-chains

The output from content structuring is fed to the Text Realisation Module, which uses the *generation-by-classification* technique (Reiter and Mellish, 1992) to assert some

facts about the text to be realised. This process recursively builds an RST tree from the entity-chain structure, making explicit the hierarchical levels of each complex span (an RST subtree) and creating a clause node for each atomic span (a fact node) in an entity-chain. There are type restrictions on each node of the tree. The RST tree is then passed to the Post-processing Module, where tasks like noun phrase content determination and aggregation are performed.

In the original ILEX, aggregation works on the partially built RST tree to combine subtrees connected by some rhetorical relations. The aggregation module can handle simple embedding and parataxis. For embedding, when a potential NP is encountered, the aggregation module looks for the corresponding entity-chain and can embed the facts with the lowest relevance in that chain into the NP. It can embed a whole entity-chain representing a paragraph, resulting in fewer paragraphs. For parataxis, the module checks all the facts connected by JOINT and combines adjacent ones with identical Arg1 and predicate, or two adjacent facts with the same Arg1.

Finally in the Presentation Module, the completed RST tree is converted into the desired format, which can be ASCII, HTML or spoken text.

As shown in Figure 7.4, ILEX uses a pipeline generation architecture, which does not allow much interaction between different modules. The space of aggregation is also very limited. To enable more interactions between aggregation and other generation tasks, we have to modify this architecture accordingly. First of all, we need abstract syntactic constraints on content structuring. Then we need to bring aggregation considerations into the structuring process. In the next section, we describe the changes we incorporated into the ILEX system, which results in a new system called ILEX-TS (ILEX based on Text Structure).

7.3.2 Resources of ILEX-TS

We implemented the improved version of Meteer's Text Structure ((Meteer, 1992; Panaget, 1997), introduced in Section 7.2) into ILEX as the intermediate level of representation between content structuring and sentence realisation. The Text Structure replaces the RST tree produced by the original ILEX text planner, providing a

similar mechanism for abstract sentence planning. It resembles an RST tree above the sentence level and also uses this same tree representation for sentence structures.

Text Structure node

A Text Structure is a tree composed of nodes and the relations among them. We use the WAG unit structure to represent information needed in a Text Structure node, which mainly contains the following attributes:

- `features`: containing subsuming ideational concepts in the Upper Model (introduced in Section 3.2.2);
- `roles`: specifying the following relations:
  - `sem`: the content of the node, which can be a fact node or an entity node in the Content Potential (introduced in Section 7.3.1);
  - `syn`: the textual semantic category of the node (introduced in Section 7.2);
  - `relation`: the relation among its children, which can be RST-ELABORATION or RST-SEQUENCE, etc. These relations stand both above and within clauses;
- `backpointers`: the relations between this node and its parents.

Overview of the generation algorithm and resources

The following simplified algorithm describes the ILEX-TS text generation procedure as a whole:

1. The content selection module selects a set of facts to be expressed, which forms a relevance queue, as in ILEX;
2. The entity-chains text planner forms entity-chains and computes the best RST trees that can be included in the chains. See (Knott and O'Donnell, 1998) for a detailed description of this procedure and the set of RST relations used in ILEX.
3. Instead of building an RST tree out of the entity-chains as in ILEX, the text structuring module of ILEX-TS checks through the entity-chains and recursively

expands the Text Structure when a new unit of a chain or a new chain is encountered. NP form determination and embedding also happen at this stage. The detailed algorithms will be given in Section 7.3.3.

4. Step 3 repeats until all facts are consumed. Then the aggregation module searches through the Text Structure for parataxis possibilities on adjacent substructures;

5. The fully built and appropriately simplified Text Structure is sent to the surface realiser, where the complete text including nominal groups is generated, as in ILEX.

Our focus is on the third step of the above algorithm, where a Text Structure is constructed out of the entity-chains. This process makes use of a number of resources, including the hierarchy of ideational semantic categories, the hierarchy of textual semantic categories, resource trees and embedding rules. We introduce these resources below and describe how they are used in the TS construction in Section 7.3.3.

Resources

a hierarchy of ideational semantic categories (ISC) : a hierarchical organisation of the concepts (i.e., things, processes, properties, etc.) that may be expressed in languages. This hierarchy is built from the WAG ontology (O'Donnell, 1994) and has three major ideational categories: Thing, Configuration and Quality. The Configuration sub-hierarchy is rewritten according to the Generalised Upper Model (Bateman et al., 1995), as described in Section 3.2.2.

a hierarchy of textual semantic categories (TSC) : a domain independent inheritance network which organises concepts according to their possible textual realisations (e.g., Time-anchored-clause, Intransitive-nucleus, Clause-modifier, etc.). It is built according to (Panaget, 1997) (described in Section 7.2). A fragment of the hierarchy is shown in Figure 7.6 (the `tsc-` prefix before each concept is omitted for simplification, except for `tsc-unit`).

For each category in the hierarchy, a `selection-constraint` role can be specified, which gives the requirement that must be satisfied in order to use this

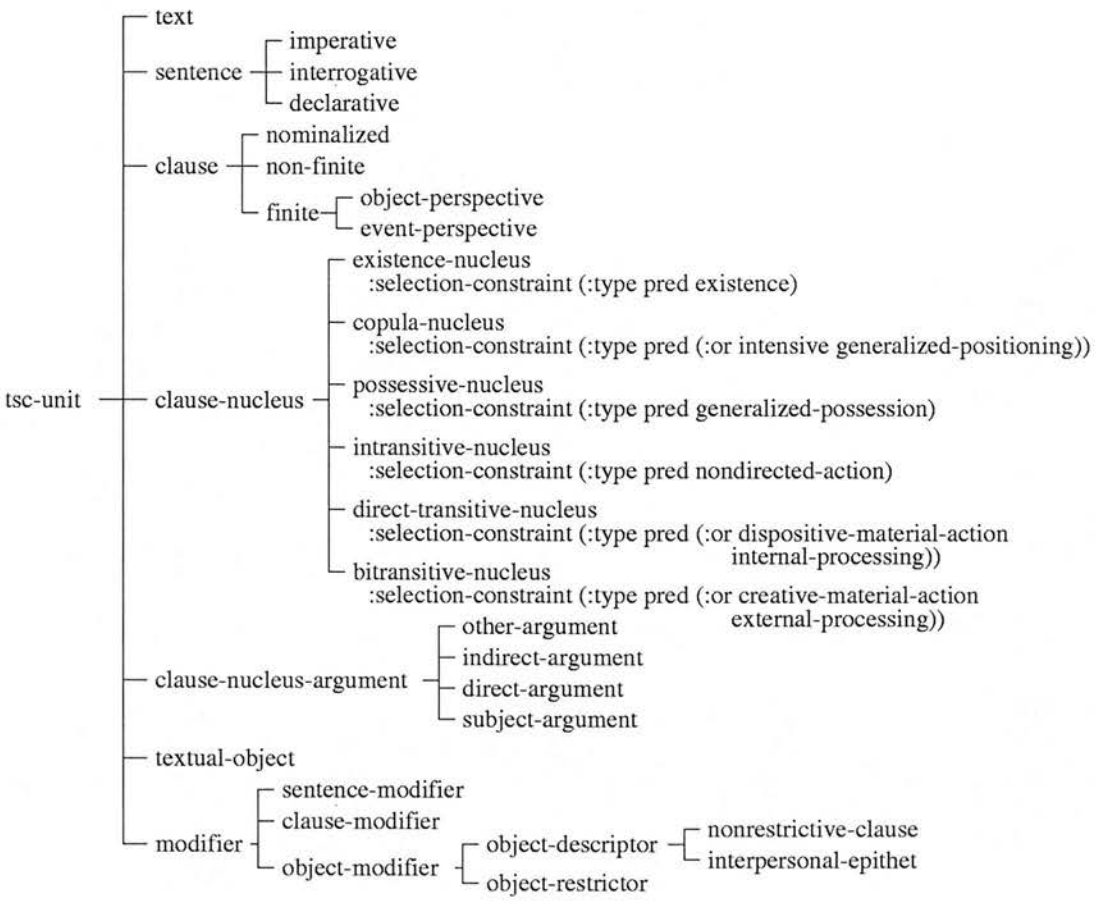


Figure 7.6: A fragment of the hierarchy of textual semantic categories

category. For example, to realise the predicate of a fact with one of the six sub-concepts of `tsc-clause-nucleus`, the predicate needs to be subsumed by the ideational concept `Creative-material-action` or `External-processing` in order for `tsc-Bitransitive-nucleus` to be chosen, or subsumed by `Nondirected-action` in order for `tsc-Intransitive-nucleus` to be chosen to capture the textual property of the predicate. If there is no constraint on the sub-concepts of a category, the last sub-concept will be selected by default.

This hierarchy is used to map from ideational semantics to textual semantics, which restricts the syntactic structure that can be used to realise an ideational concept. This mechanism resembles the chooser of Penman (Mann and Matthiessen, 1985).

15 resource trees : predefined structures corresponding to textual semantic cate-

gories for building and expanding the Text Structure. In the following example, `name` specifies the textual semantic category for which this resource tree is defined. So if the `syn` role of the current TS node has the value `tsc-possessive-nucleus`, the tree *RT* below can be used to expand the node; `rel-child` gives the relation of the current TS node with respect to its children when this resource tree is used; `syn-func` specifies the function for building the concrete syntactic structure of the named textual semantic category (this is specific to ILEX); `restrict` is the restrictions on the children nodes, which is itself a complex structure. `cardinality` specifies how many times this child should be presented. Other restrictions on a child node include the relation to its parent node, its textual semantic category and its ideational semantic category.

$$
\left[\begin{array}{l}
\text{name : tsc} - \text{Possessive} - \text{nucleus} \\
\text{rel} - \text{child : Head} \\
\text{syn} - \text{func : nil} \\
\text{restrict :} \left[\begin{array}{ll}
\text{c1 :} & \left[\begin{array}{l}
\text{rel} - \text{parent : Argument} \\
\text{textual} - \text{sem : tsc} - \text{Subject} - \text{argument} \\
\text{ideational} - \text{sem : Possessor} \\
\text{cardinality : one}
\end{array}\right] \\
\text{c2 :} & \left[\begin{array}{l}
\text{rel} - \text{parent : Argument} \\
\text{textual} - \text{sem : tsc} - \text{Direct} - \text{argument} \\
\text{ideational} - \text{sem : Possessed} \\
\text{cardinality : one}
\end{array}\right]
\end{array}\right]
\end{array}\right]_{RT}
$$

All the resource trees are stored in a hash table, with `name` as the key. When a resource tree is not defined for a category, the tree of a subsuming concept will be used for expanding the Text Structure. The set of resource trees provides the abstract syntactic restrictions on document structuring.

74 mapping relations between predicates and ideational concepts as well as syntax : the mapping from the predicate of a fact to an ideational semantic category and other syntactic features. This establishes a connection between the predicate of a fact, the subsuming ideational concept and the way the predicate

will be expressed (this is more concrete than the TSC categories and is specific to ILEX), for example,

$$\begin{bmatrix} \text{entity} - \text{type} : \text{jewellery} \\ \text{role} : \text{designer} \\ \text{process} : \text{Creative} - \text{Material} - \text{Action} \\ \text{tense} : \text{past} \\ \text{voice} : \text{passive} \\ \text{verb} : \text{make} - \text{verb} \end{bmatrix}_{Map}$$

`entity-type` gives the class of Arg1 fillers, which if specified can be used to override a general case expression when more than one possible expression is provided for a predicate; `role` gives the name of the predicate whose expression is being defined; `process` specifies the ideational concept subsuming the predicate; and `verb` is the lexical item used to realise this predicate. The meaning of the other attributes should be clear from their names.

All the mappings from predicates to ideational categories are stored in a hash table, with the list consisting of `role` and `entity-type` as the key.

6 embedding rules : rules for embedding (described in Section 4.2.2) on the Text Structure. Each rule enables one type of embedding to be performed during content structuring.

In the following example, `priority` gives the order in which the rule should be tried, where those rules producing simpler syntactic forms always have higher priority (Scott and de Souza, 1990); `type` specifies the type of embedding; `constraints` are the conditions that must be satisfied by the predicate and arguments of the fact to apply this rule and by the realisation of the referring part; and `RT` specifies the restrictions on the resource tree for expressing this modifier. The tree satisfying the restrictions can be used to expand the current TS node.

$$\left[\begin{array}{l} \text{name : eval} - \text{premod} \\ \text{priority : 5} \\ \text{type : epithet} \\ \text{constraints :} \left[\begin{array}{l} \text{pred : Property} - \text{Ascription} \\ \text{arg2 : Evaluative} - \text{quality} \end{array}\right] \\ \text{RT :} \left[\begin{array}{l} \text{rel} - \text{parent : Adjunct} \\ \text{textual} - \text{sem : tsc} - \text{Interpersonal} - \text{epithet} \end{array}\right] \end{array}\right]_{Rule}$$

Although the decision tree given in Section 4.3 presents more fine-grained embedding rules than those in Section 4.2, we do not use it in ILEX-TS because ILEX only has very limited types of expression and the rules in Section 4.2 are sufficient. ILEX-TS currently has six embedding rules, which reflect the rules summarised in Table 4.2. They cover the generation of evaluative, quality and class adjectives, prepositional phrases signalled by "with" and non-restrictive relative clauses. They are stored in an array, ordered by their priorities. The small number of rules is also due to the restricted types of expression available in ILEX. In other domains, more embedding rules might have to be used.

All the above resources are built using the functions provided by WAG.

7.3.3 Building the Text Structure

We mentioned in the previous section that the text structuring module built the Text Structure from entity-chains in place of an RST tree. This section describes the TS construction and embedding algorithm in detail.

Algorithm for building the Text Structure

Taking a sequence of ordered entity-chains as the input, the TS construction is a recursive process working as follows:

1. For each text span in an entity-chain, the current and previous *Cb* are updated, i.e., assigning the current *Cb* to the previous *Cb* and the Arg1 of the main fact of the span to the current *Cb* (we assume that the *Cb* is usually the Arg1 of a fact).

- If the span is a complex one, i.e., an RST subtree, the algorithm proceeds depth-first, constructing a TS node for each top-level relation. Suppose the tree is Rel(Span1,Span2), which means that the relation Rel connects the text spans Span1 and Span2. One of the following four types of TS sub-structure will be built for the tree, depending on the rhetorical relation being used:
  - Span1 is the parent node and Span2 an adjunct node of type `tsc-clause-modifier`, for relations like RST-SIMILARITY;
  - A new TS node is created as the parent node with Span1 as the matrix child and Span2 an adjunct child, for relations like RST-CONCESSION, realisable as a single sentence with a main clause and a subordinate clause;
  - A new TS node is created as the parent node with Span1 and Span2 as two coordinated children, for relations like RST-WHEREAS, realisable as a coordinated sentence;
  - A new TS node is created as the parent node with Span1 as the matrix child and Span2 an adjunct child, for relations like RST-AMPLIFICATION, realisable as a complex sentence.

  The `relation` role of each top-level node is specified as the relation in the RST subtree.

- If the span is an atomic one, i.e., a single fact, the TS for the fact is built, making use of resource trees. The constructor first builds a TS node whose `syn` is `tsc-clause`, which requires exactly one matrix child and any number of adjunct children. There is no restriction on the ISC value at this point. Starting with this simple restriction, i.e., a general TSC, the algorithm retrieves the ISC from the predicate of the fact, using the mapping from predicates to ideational concepts. Given the general TSC and the ISC, a more specific TSC can be retrieved from the Hierarchy of TSC. The resource tree of the specific TSC or the more general one can then be used to expand the current TS node. The tree also posses constraints on the children nodes that can be used for TS expansion.

This recursive process starts building TS node whose `sem` is a fact node, then proceeds to build TS nodes for the predicate, Arg1 and Arg2 of the fact node. It continues until there is no more TS node to expand.

2. Between text spans inside an entity-chain, as well as between entity-chains, the algorithm builds TS nodes whose `relation` slots are specified as RST-JOINT to connect text spans and entity-chains.

The process for complex text spans is a slight revision of the original ILEX process for text realisation. The description is simplified as the actual process also considers the level of an RST node and nuc/sat ordering (more details can be found in (Knott and O'Donnell, 1998)). The process for complex text spans does not use resource trees for building the TS, simply because we tried to make use of the existing ILEX code as much as possible.

Algorithm for NP content determination

The previously proposed methods for generating complex referring expressions are not capable of capturing the complex interaction between RE components. Most of them perform aggregation before the referring process, e.g., (McKeown et al., 1997; O'Donnell et al., 1998), so they do not consider the effect of embedding on local coherence and possible conflicts between the addition of referring and non-referring information. However, if we assume that embedding comes after, then it cannot affect the choice of NP forms. So either way the interaction cannot be captured satisfactorily and we argue that the two processes should be more intertwined.

In Section 3.5 we mentioned that to take into account various considerations like coherence and style, it is highly possible for an entity to have multiple realisations and psycholinguistic research has shown that people cannot reach a consensus on the syntactic realisation of an entity in many cases, e.g., (Yeh and Mellish, 1997). Our algorithm is based on the method described in (O'Donnell et al., 1998) and it allows multiple realisations for a discourse entity, which have no significant difference in coherence. The algorithm chooses a realisation that is most suitable for embedding new information. This way, it takes into account the coherence between adjacent utterances and

embedding considerations simultaneously in the choice of NP forms.

Since the generation of the referring part is not the central issue in our research, the algorithm is simplified and is by no means to be taken as a general model for all referring problems. We use Centering Theory as an example for modelling local coherence and it can be substituted by any other theory.

The algorithm assumes two global variables, which store all focal objects by means of discourse or immediate situation, and entities that are supposed to be known to the reader. It can be described as:

1. Try all possible syntactic realisations for an entity, including

   - Pronoun: if the entity is the *Cb* of the current utterance, and the current *Cb* is the same as the *Cb* of the previous utterance;
   - Full name: if the entity is a discourse new referent and has a full name;
   - Short name: if the entity is a discourse old referent and has a short name;
   - Demonstrative: if the entity is one of the focal objects, but not the *Cb* of the previous utterance;
   - Definite: if the entity is one of the shared entities, or if the entity is somehow associated with a shared entity, or if the appositive construction *definite NP(,) PN* is possible (i.e., satisfying the conditions of the algorithm in Section 4.3.4). Possessive definite belongs to this category;
   - Indefinite: if none of the above applies.

   All suitable realisations are stored in a list *NP-forms*.

2. Compare the realisations in *NP-forms* with that for the same entity in the previous utterance (if the entity is mentioned there) and remove duplicates from the list. More precisely, if the same pronoun, demonstrative, definite or short name form for the same entity is used in the previous utterance, remove it from *NP-forms* unless it is the only choice.

3. Order *NP-forms* according to: *Pronoun* $\prec$ *Demonstrative* $\prec$ *Full-name* $\prec$ *Definite* $\prec$ *Short-name* $\prec$ *Indefinite*, where A $\prec$ B means that A is in front of B. The ones

located in the front are better choices.

4. If *Pronoun* is not the only member of *NP-forms*, the embedding process will make decisions based on the constraints imposed by the first form in *NP-forms* except for *Pronoun*, which include what to embed and which syntactic form the embedded parts should use.

5. If there is embedded information, the first form in *NP-forms* except *Pronoun* is the best realisation, otherwise the first of *NP-forms* is chosen.

6. If *Definite* is finally chosen as the realisation, check the properties and syntactic slots that are needed for referring. This process can make use of the incremental algorithm of (Dale and Reiter, 1994) and the preferences for semantic features given in Section 4.3.3. If a slot has already been occupied by the embedded information, free the slot by removing the embedded part.

In this algorithm, the goals of referring and maintaining local coherence have higher priority over the goal of expressing more information, which is only satisfied when it does not interfere with the more important ones. The algorithm has the advantage of being simple and easy to be incorporated into existing NLG systems. It gives one way to capture the interaction, but we do not claim that it is the best way.

NP content determination in ILEX-TS

In ILEX-TS, NP content determination happens in the TS construction process rather than in post-processing as in ILEX. After the TS for each atomic text span is built, the algorithm checks that span for possible embeddings. It works like this:

1. For the Arg1 and Arg2 of a fact whose TS has just been built, the algorithm finds all possible realisations of the entities as described above.

2. For the two entities, all facts in the corresponding entity-chains are collected into two lists, except for complex spans in the chains because these are interesting information in our domain and should not be realised as embedded components.

3. If a list is not empty, it is ordered with respect to the relevance of the facts. The algorithm starts with the least relevant fact and checks if it satisfies the constraints of an embedding rule and if there is an available syntactic slot around the head (to be clarified in Section 7.3.4). When a rule is found, other constraints are also checked, including:

   - whether the Arg2 of the to be embedded fact is the topic of a later entity-chain, in which case the fact should not be embedded because embedding will prevent the smooth introduction of the entity-chain into the context.
   - whether hypotaxis or semantic parataxis is possible between the to be embedded fact and some other facts. If so, it should not be embedded.

   Due to the complexity of the cases that need to be considered, some other hand tuning might be necessary. For a fact whose predicate is "isa", embedding only happens in the Arg2.

4. The most appropriate syntactic forms of the entities are decided, as well as the properties needed for unique identification. If there is a fact to be embedded, the TS for the embedded component is built using the restrictions from the RT slot of the embedding rule. The chosen facts are then deleted from the corresponding entity-chain.

We set the maximum relevance value (0-1) for embedding to 0.9. A fact with a higher value should be realised as a main clause rather than an embedded component.

The statement of the algorithm shows that the embedding decision takes into account such factors as: the syntactic form of the NP head, the preconditions of the embedding rules, the relevance of the embedded content, the availability of the required syntactic slots and the possible effect of the embedding on topic moves and other types of aggregation.

After the TS is fully built, the aggregation algorithm searches through the TS and combines sentence level text structures with two identical components among Pred, Arg1 and Arg2, which should have been positioned next to each other by the text

planner, as well as two adjacent structures with the same Arg1. This results in an overall simpler TS with more complex phrase structures.

A generated text

We give a description of a jewel generated by ILEX-TS, with the embedded parts in italics. Generally the texts produced by ILEX-TS are not very different from those by ILEX because ILEX-TS uses the ILEX text planner and realiser, which can only produce restricted discourse and syntactic structures. Therefore, ILEX-TS does not perform substantially more embedding than ILEX does and the generated text lacks variation even with aggregation.

> This jewel is a bracelet, *which is 0.6 cm in width.* It was made by a *famous young English* designer called Gerda Flockinger. The jewel has a *swelling* midsection and a *slightly flared* band. It is in the Organic style. The jewel, *which was made in London,* was made in 1968. It is made from silver metal, turquoises, tourmalines, aquamarines and pearl. The jewel draws on natural themes for inspiration, in that it is a remarkably fluid piece; indeed Organic style jewels usually draw on natural themes for inspiration.
>
> Gerda Flockinger is a designer, *who lived in London.* She was one of the best jewellers working in this medium. She got very sophisticated colour control here.
>
> Organic style jewels are usually encrusted with gems and made up of asymmetrical shapes. They usually have a coarse texture.

7.3.4 Capturing the Rules and Preferences in ILEX-TS

This section discusses how the rules defined in Section 3.5 (Rules 3.1 and 3.2) and the preferences abstracted in Section 6.2.3 (Heuristics 6.1 to 6.4) are captured in ILEX-TS.

Conforming to the restrictions on the non-referring part

For Rule 3.1, ILEX-TS uses embedding rules to prevent confusing the reader about the referent indicated by the referring part. This is mainly achieved by checking a fact and a potential list of NP forms against the `constraints` slot of a rule and only when all the specified constraints are satisfied can the fact be a candidate for embedding. That is, the constraints on embedding rules have as one of their goals to guard against confusion.

To give an example of how an embedding rule is used, assume we have two facts *F1=date(ajewel, 1905)* and *F2=hasqual(ajewel, floral-motifs)* as in (7.2a), where *F1* is the fact whose Text Structure is being built. The referring form of *ajewel* can be *Demonstrative, Definite* or *Pronoun*. Since *F2* satisfies the embedding rule *Rule1* below, which requires the predicate of the fact to be subsumed by the ideational category of `generalised-possession` and the referring head to be demonstrative or indefinite, the rule can be applied. If other checks also succeed, the algorithm will realise *F2* as a post-modifier of the *Arg1* of *F1*. Finally the Demonstrative realisation is chosen, as in (7.2b), and the resource tree of `tsc-prep-phrase` will be used for building the TS of this modifier.

(7.2) a. *This necklace was made in 1905. It has floral motifs.*

b. *This necklace* ***with floral motifs*** *was made in 1905.*

$$\left[\begin{array}{l} \text{name : with} - \text{phrase} \\ \text{priority : 4} \\ \text{type : prep} - \text{phrase} \\ \text{constraints :} \left[\begin{array}{l} \text{pred : Generalized} - \text{Possession} \\ \text{refer : this} - \text{definite/indefinite} \end{array}\right] \\ \text{RT :} \left[\begin{array}{l} \text{rel} - \text{parent : Adjunct} \\ \text{textual} - \text{sem : tsc} - \text{Prep} - \text{phrase} \end{array}\right] \end{array}\right]_{Rule1}$$

Concerning the readability of an NP (Rule 3.2), we use a similar method to what is described in (Horacek, 1997). That is, we restrict the number of modifiers that can appear around the head and only allow one level of clause embedding. A similar

method is also used in (Robin, 1993), where the addition of optional information takes into account such surface constraints as number of words and depth of embedding, etc. We allow two epithets and one classifier before the head, and two qualifiers, which can be prepositional phrases and relative clauses, after the head. However, this restriction can be relaxed depending on the domain. For example, in summaries of basketball games, very complex NP structures are usually used (Robin, 1994b).

The fixed number of slots only restricts the maximum amount of information that can be expressed. As discussed in Section 3.5, the user model decides the actual complexity of a nominal group. At present we only distinguish between adults and children. For adults, there are no extra restrictions on the amount of additional information. But for children, because non-restrictive clauses in subjects are the major factors reducing readability, no such clauses are generated. We use a global variable to store user configurations and the variable is checked by the embedding algorithm.

Modelling the constraints among coherence features

Heuristics 6.1, 6.2 and 6.3 are followed naturally by the entity-chains text planner of ILEX, which calculates the best RST trees and puts facts connected by the imaginary CONJUNCT relation next to each other. If a fact is consumed by an RST relation other than CONJUNCT, JOINT or ELABORATION, it cannot appear directly in other relations. So expressing rhetorical relations has priority in consuming unused facts over other processes. The Text Structure built from entity-chains inherits these properties. Since parataxis is only performed on adjacent facts that have at least two identical parallel parts, connected by the ELABORATION or JOINT relation, and are at the same level of the RST tree, it is not possible to perform undesirable parataxis on an RST subtree connected by a semantic relation.

Due to the way entity-chains are formed (see Section 7.3.1), center continuations appear naturally between text spans inside an entity-chain. Between chains, a resumption relation connects an entity-chain with a discourse entity mentioned in a previous chain. The planner does not handle associate shifting at the moment. So this planning strategy maximises continuation but emphasises less on other types of center transition.

The embedding rules with different priorities favour a good embedding over a normal one. A bad embedding is not allowed at all. In order for different subtypes of aggregation to be coordinated among themselves, for each nucleus fact and the fact to be embedded, the embedding algorithm checks their parataxis and hypotaxis possibilities. Embedding is not allowed if the embedded properties cannot be realised as a syntactic form other than a non-restrictive clause in paratactic nuclei, or if not all of the paratactic or hypotactic facts of the fact to be embedded can be embedded at the same time. The preference about the combination of embedding and center transitions is captured by the first point of the third step of the NP content determination process (Section 7.3.3). These realise Heuristic 6.4.

7.3.5 Summary and Discussion

Section 7.3 describes how the rules and preferences introduced in Chapters 3 and 6 are implemented in ILEX-TS, an object description generation system using a slightly modified pipeline architecture. The NP generation algorithm reflects the bilateral relation between the referring and the non-referring part to some extent and it enables ILEX-TS to produce nominal groups serving multiple functions. Most of the preferences among coherence features can be captured in a simple way by the ILEX text planner and the aggregation processes.

However, there are a few problems with this architecture:

1. The heuristics are only intended to give general guidelines for text planning and not to be taken as hard requirements. In fact, many relations between generation related factors are better treated as preferences, which is hard for a pipeline architecture to deal with. Besides, in ILEX-TS, a feature is either present or not present, and the gradual differences in preferences cannot be modelled.

2. Although in ILEX-TS, embedding happens earlier than that in ILEX, it comes after entity-chain construction and therefore has to make decisions under the constraints of existing coherence. Embedding is still not incorporated in the text planning.

3. Although ILEX-TS can model most of the preferences we have mentioned, the

various features are optimised in order rather than simultaneously and no backtracking is performed, so there is no guarantee that the best overall text will be found.

4. A pipeline architecture allows few alternations and limits the space for exploring good unknown text properties. A better generation architecture should allow equally good texts with different structures to be produced for us to study how factors contribute together to a coherent text.

The above problems exist for a pipeline architecture no matter whether the text planner uses a top-down or a bottom-up strategy. To fix them, we need an unconventional architecture for modelling aggregation. In the next section, we describe such an architecture which can capture the interactions between tasks more easily.

7.4 Aggregation in GA-plan

We choose the text planner using a Genetic Algorithm (GA) as proposed in (Mellish et al., 1998a) for our second implementation. We call this system GA-plan. In this section, we describe this experimental system in detail, with an emphasis on how it models the interactions between generation tasks.

7.4.1 Why GA?

If we treat text generation as a search problem, many search methods can be used. However, exhaustive search or a constraint-based approach (Marcu, 1997b) is not suitable because the search space is too big. For one example text containing only 7 facts, the number of possible orders is $7! = 5040$, and if considering all embedding possibilities, the number of combinations for this specific text is 322,560. One advantage of a GA is that it can sometimes find the right track through a large space fairly quickly. However, there is no rigorous answer to when GA is a good method to use.

In the GA community, many researchers share certain intuitions for applying GA, which seem to be compatible with the properties of natural language generation. Mitchell (1996) states that a GA will have a good chance of being competitive with or

surpassing other general-purpose methods which do not use domain-specific knowledge in their search procedure when

- *the search space is large* so that exhaustively searching for the best solution is impractical: a normal text in our domain contains at least 15 facts, which produces an incredibly huge number of combinations. Some texts do have many more facts.
- *the search space is known not to be perfectly smooth and unimodal* (i.e., does not consist of a single smooth "hill") so that a gradient-ascent algorithm like steepest-ascent hill climbing is less efficient: it is generally agreed that there could be many ways to express pieces of information. The differences may come from ordering, structuring or expressing, etc. and different expressions can be equally coherent. It is not clear whether there is an optimal way of expressing information.
- *the search space is not well understood* so that search methods using domain-specific heuristics cannot be devised: although a great amount of research has been done in natural language generation and people have gained some knowledge about it, many questions remain unanswered due to the complexity of the problem.
- *the task does not require a global optimum to be found*, i.e., if quickly finding a sufficiently good solution is enough: people are not bothered with expressing information in the optimal way because it is not necessary for effectively exchanging ideas among communication agents. Generally a sufficiently coherent text would be enough.

In general, a GA approach might provide a suitable mechanism for the problem of natural language generation, or at least some aspects of it.

7.4.2 The Problem and the Input

The problem of given a set of facts and a set of semantic relations between the facts, producing a legal RST tree using all the facts and some relations was first charac-

terised by Marcu (1997b), where a constraint satisfaction approach was used to target the problem. Mellish et al. (1998a) further develops the idea by experimenting with different stochastic search methods and the genetic algorithm they use seems to produce reasonably good texts.

In this problem, the input is a set of facts and a set of relations between them, a fragment of which looks like that in Figure 7.7.

```
top_focus(choker).
fact(choker,be,broad,fact_node-201,0.6).
fact(choker,'is fitted','closely around the neck like a dog-collar',fact_node-
    202,0.9).
fact('Queen Alexandra',wore,choker,fact_node-203,0.9).
fact(choker,'can cover',scar,fact_node-204,0.9).
fact(scar,be,small,fact_node-205,0.5).
fact(scar,be,'on her neck',fact_node-206,0.7).
fact(band,'might be made of',panels,fact_node-207,0.8).
fact(panels,is,'discreetly hinged',fact_node-208,0.8).
fact(band,'might be made of',plaques,fact_node-209,0.8).
...
rel(in_that_reln,fact_node-203,fact_node-204,[],0).
rel(disjunct,fact_node-207,fact_node-209,[],0).
```

Figure 7.7: A fragment of the input to the GA text planner

top_focus indicates the specific object the text is to describe. Each fact is represented in terms of a subject, a verb and a complement, as well as a unique identifier and the interestingness/relevance of the fact. Each relation is represented in terms of the relation name, the two facts (nucleus/satellite or multi-nuclear) that are connected by the relation and a list of precondition facts which need to have been mentioned before the relation can be used. In GA-plan, relations have an extra slot for *inferrability* (described in Chapter 5), which is set to 1 or 0 for strong or weak inferrability respectively. This is to model the psycholinguistic observations summarised in Section 5.3.4, i.e., embedding might be a good alternative expression for a semantic relation between two facts if the relation is strongly inferrable. Such embeddings are preferred over those that do not have any semantic connection in between, which could be one way of controlling the relevance between the main clause and the NR-clause. Generally,

inferrability has to be implemented based on limited domain-dependent knowledge and user configuration.

As GA-plan is an experimental system, its ability is limited in many aspects. It does not have a real realisation component, so the parts we are less interested in are represented by canned phrases for readability. However, it should be understood that the canned phrases correspond one to one with entities in the museum domain.

Problem encoding

In early GAs, binary encodings of the target problems are the most common encodings and much of existing GA theory is based on the assumption of fixed-length, fixed-order binary encodings (Mitchell, 1996). However, they are unnatural and unwieldy for many problems. Davis (1991) strongly suggests using whatever encoding that is most natural for the actual problem, and then devising a GA that can use that encoding. This philosophy is adopted by much current research.

Mellish et al. (1998a) compare ordinal encoding and path encoding and adopt the latter because it is more natural. Path encoding represents a candidate solution as a sequence of facts. For example, a path encoding of the facts in Figure 7.7 is [fact_node-204, fact_node-201, fact_node-209, fact_node-207, fact_node-203, fact_node-202, fact_node-205, fact_node-206, fact_node-208]. An RST tree can be built deterministically from such a sequence (to be elaborated in the next section) and the tree can then be realised as a text.

7.4.3 The Planning Procedure

Given the sequences as the input, the GA-based text planning is basically a repeated two step process - firstly sequences of facts are generated by applying GA operators and secondly the RST trees built from these sequences are evaluated. (Mellish et al., 1998a) summarises the genetic algorithm roughly as follows:

1. Enumerate a set of random initial sequences by loosely following sequences of facts where consecutive facts mention the same entity.

2. Evaluate sequences by evaluating the rhetorical structure trees they give rise to.

3. Perform mutation and crossover on the sequences, with mutation having a relatively small probability.

4. When the "best" sequence has not changed for a time, invoke mutation repeatedly until it does.

5. Stop after a given number of iterations, and return the tree for the "best" sequence.

In the algorithm, the rhetorical structure trees are right-branching and are almost deterministically built from sequences of facts. The algorithm always uses a normal semantic relation if there is one between two facts, otherwise, it uses a CONJUNCT or DISJUNCT relation (as described in Section 6.1.3); when all these fail, it uses a JOINT relation.

The advantage of this approach is that it provides a mechanism to take into account various planning preferences in the fitness function (to be described in Section 7.4.6) and search for a rhetorical structure tree featuring the best combinations of coherence properties at a given moment.

7.4.4 GA Operators

Intuitively, we know that the ordering and adjacency of information affect coherence. The devised GA operators should be able to maintain much of the desired ordering and adjacency from an old generation to a new one. Mellish et al. (1998a) propose two operators: a crossover and a mutation, which create new sequences from an existing population.

To explain the GA operators, we introduce a *unit* structure, which can be either a fact or a list of facts or units with no length limit. A sequence is composed of units. For a complex unit (i.e., a list) in a sequence, we call its very first fact the *main fact*, into which the remaining facts in the list are to be embedded. For a unit which is a single fact, the fact itself is the main fact.

Given two sequences, the crossover operator performs a specific two point crossover, which inserts a random segment from one sequence into a random position in the other to produce a new sequence. For example, given two sequences,

$[U_{11},...,U_{1i},...,U_{1j},...,U_{1n}]$ and $[U_{21},...,U_{2k},...,U_{2m},...,U_{2n}]$

the crossover first selects a segment from a sequence, say $U_{1i},...,U_{1j}$ from the first one, then selects a random position from the other sequence, say *2m*, and inserts the segment into the selected position. The duplicates outside the inserted segment are removed. This produces a new sequence,

$[U_{21},...,U_{2m-1},U_{1i},...U_{1j},U_{2m},...,U_{2n}]$

The mutation operator selects a random unit of a sequence and moves it into a random position in the same sequence. For example, given a sequence,

$[U_1,...,U_g,U_i,...,U_j,...,U_n]$

the mutation randomly selects two positions, say g and j, and moves U_g to before j to produce a new sequence,

$[U_1,...,U_i,...,U_g,U_j,...,U_n]$

We call this operation the *normal mutation.*

To explore the whole space of aggregation, we decided not to just perform aggregation on rhetorical structure trees or on adjacent facts in a linear sequence because they might restrict aggregation possibilities and even miss out good candidates. This is similar to the work of (Shaw and McKeown, 1997; Shaw, 1998a) to the extent that their algorithms also search through sequences of clause-sized semantic representations for possible aggregations rather than only combining adjacent ones.

To maximally explore aggregation, we define a third operator called *embedding mutation.* The embedding mutation randomly selects a unit from a sequence and an entity in its main fact. It then collects all the units outside the selected unit whose main facts mention this entity, and randomly chooses one. The list containing these two units represents a random embedding and will be treated as a single unit in later operations. It takes the position of the first original unit and those two units are removed from the

sequence. This produces a new sequence, which is then evaluated and ordered in the population. For example, given a sequence,

$[U_1,...,U_g,U_i,...,U_j,U_k,...,U_n]$

the embedding mutation randomly selects a unit, say U_i, and a possible embedding, say U_k, and has $[U_i,U_k]$ as a new unit in the position of U_i. The new sequence is,

$[U_1,...,U_g,[U_i,U_k],...,U_j,...,U_n]$

and repetitions outside $[U_i,U_k]$ are removed. Note that U_i and U_k can be simple or complex units.

Normally GA operators would not change the length of a solution, but researchers are exploring different types of operators, such as messy GAs, which build up increasingly longer, highly fit strings from shorter blocks. In our case, the length of a solution is not an important property, and shorter sequences usually contain more desirable properties.

One issue we have not mentioned is how to select the individuals in a population to create offspring. The purpose of selection is to emphasise fitter individuals and expect their offspring to have even higher fitness. However, selection has to consider both diversity and evolution speed. (Mellish et al., 1998a) uses rank selection, which has a bias towards the fittest elements. The probability of selecting a given element is R times the probability of selecting the previous one in the sequence of elements sorted by fitness. R needs to be a number less than 1, but if it is too small, there will be a negligible chance of picking elements with very low fitness. A good default for R is 0.95.

7.4.5 Parameters for the Genetic Algorithm

The values for GA parameters like population size, crossover rate and mutation rate can affect the performance of a GA system significantly. "These parameters typically interact with one another nonlinearly, so they cannot be optimised one at a time. There is a great deal of discussion of parameter settings and approaches to parameter adaptation in the evolutionary computation literature ... There are no conclusive

results on what is best; most people use what has worked well in previously reported cases." (Mitchell, 1996)

One well-known piece of work in parameterisation is De Jong's study of genetic algorithms in function optimisation. De Jong (1975) performed a series of parametric studies across a five-function suite of problems and suggested that good GA performance requires the choice of a high crossover probability, a low mutation probability (inversely proportional to the population size), and a moderate population size (50 - 100). According to this study, smaller populations have the ability to change more rapidly and thus exhibit better initial on-line performance. Schaffer et al. (1989) systematically tested a wide range of parameter combinations and suggested population size 20-30, crossover rate 0.75-0.95, and mutation rate 0.005-0.01. An experiment in (Goldberg, 1989) also used the population size 30, and the probabilities for crossover and mutation 60% and 3.33% respectively.

The reason for a high crossover probability is that crossover combines the good bits of two sequences and is more likely to produce a new sequence bearing desired properties. In contrast, mutation is entirely random. In our case, the crossover is likely to result in a new RST tree combining two good sub-trees (as each segment of a sequence corresponds to an RST sub-tree), whereas the normal mutation is unlikely to have such an effect. The probability of applying the embedding mutation should be reasonably high as this allows certain amount of embedding, which is a desirable property of a coherent text.

We follow the general methodology and set the population size to 30. Despite the good qualities of crossover, it does have drawbacks. Since crossover copies segments from mainly successful individuals and spreads these segments around, if the successful individuals do not happen to have certain properties in them, these properties will not get spread and will probably disappear from the population. Therefore, a crossover might cause a small population to "lose" information in such a way that the algorithm cannot make use of it again unless it is accidentally reintroduced via mutation. For this reason, we increase the probability of using the normal mutation to 4% and apply the crossover and the embedding mutation with the probability of 65% and 31% respectively.

7.4.6 The Evaluation Function

A key requirement of a GA approach is the ability to evaluate the quality of a candidate solution. This is the task of the fitness function of the algorithm. In the NLG context, we will call it the evaluation function because it is mainly used to evaluate the coherence of a solution text. In the specific GA proposed in (Mellish et al., 1998a), the quality of a sequence is indicated by the overall score of the rhetorical structure tree the sequence gives rise to, which is the sum of positive and negative scores for all the good and bad properties the tree bears. Those sequences scored higher are kept for producing better offspring.

Their scheme scores some basic features of an RST tree, for example, +21 for a semantic relation other than JOINT and ELABORATION, which prefers the use of more interesting relations, and -9 for a fact (apart from the first) not mentioning any previously mentioned entity, which prefers a smooth center transition over an abrupt one. However, they make it clear that the scores are there for descriptive purposes rather than for making any serious claim about the best way of evaluating RST trees.

There has been much linguistic and psycholinguistic evidence of preferred properties of text according to human authors, some of which are captured by the heuristics in Chapter 6. However, these only give evidence in qualitative terms. For a GA-based planner to work, we have to come up with actual numbers that can be used to evaluate a tree.

Based on the claim in Chapter 6 that it is the relative weight among coherence features that decides the quality of the generated text, rather than the weight for each feature, we will not argue for the numbers used here except that they satisfy all the preferences mentioned in Section 6.2.3. In Chapter 8, we will show that these numbers do capture some truth about the notion of a coherent text and different sets of numbers satisfying the same preferences will agree with each other on the relative quality of texts.

Our scheme is an extension to the scoring scheme of (Mellish et al., 1998a), with the addition of scores for features of aggregation and additional types of center transition. The only extra criterion for weight assignment is that negative numbers are given to unfavourable features and positive numbers to favourable ones. Sometimes large

negative numbers are used to prevent very bad features from appearing in the solutions.

The scores for using semantic relations are (we mentioned in Section 6.2 that we do not consider ELABORATION as an explicit relation):

+21 for a relation other than JOINT, CONJUNCT or DISJUNCT.

+15 for a CONJUNCT or DISJUNCT relation, not inside other semantic relations.

-20 for a JOINT relation.

-50 for a CONJUNCT relation inside other semantic relations, a consecutive use of the same semantic relation or an embedding in a semantic relation.

Suppose we have two facts *Fact1=fact(O1,P1,O2)* and *Fact2=fact(O11,P11,O22)* next to each other, the scores for center transitions are:

+20 for a continuation: O1 = O11 (we assume Arg1 = *Cb*).

+16 for an associate shifting: there is an association relation between an object in *Fact1* and an object in *Fact2*, or two objects in *Fact1* and *Fact2* have association relations to the same object.

+14 for a smooth shifting: O1 $\neq$ O11, O2 = O11.

+6 for a resumption of a previous focus: *Fact2* mentions an entity not in *Fact1* but in the previous discourse (Knott et al., in press).

Center transitions are considered between every two adjacent facts. In the worst case, two facts are only connected by a JOINT. The above scheme will reward those JOINTs that could have been described as OBJECT-ATTRIBUTE ELABORATIONs as they are not always incoherent. In our domain, focus retaining seldom appears, so we do not model it here.

The scores for embeddings are (see Section 6.2.2 for features of different types of embedding):

+10 for a good embedding in a discourse new reference.

+6 for a good embedding in a discourse old reference or a normal embedding in a discourse new reference.

+4 for a normal embedding in a discourse old reference.

-30 for a bad embedding.

Scores for other features considered in (Mellish et al., 1998a) include:

-10 for a top nucleus not mentioning the topic of the text.

-4 for each fact that will come textually between a satellite and its nucleus.

-30 for an unsatisfied precondition for a relation.

+8 for the first fact with a given entity as subject having verb "is".

(Mann and Moore, 1981) also uses preference rules with numbers to evaluate the quality of a protosentence in order to choose the aggregation rule that results in the largest gain. The factors they considered are limited and their numbers are rather arbitrary.

7.4.7 Other Components of GA-plan

In addition to the main properties introduced in the previous sections, GA-plan has a few other components that make it a relatively complete generation system.

Decisions about the referring part

The NP form determination algorithm follows that described in Section 7.3.3, i.e., is similar to the algorithm of ILEX-TS, except that it does not try to avoid repetitive uses of the same syntactic form. All NP forms chosen by the referring process are stored in a global stack in the form *[Fact, Core1, Core2]*, where *Core1* and *Core2* represent the referring parts for the Arg1 and Arg2 of *Fact*. The referring process makes decisions according to such knowledge as the previous *Cb*, the current *Cb* and entities assumed to be a part of the shared knowledge of the hearer.

Since the referring process is not our major concern, we have not tried to incorporate the evaluation of a referring part into the GA framework. The referring parts are decided while the RST tree of a sequence of facts is being constructed.

The decision as to whether an embedding is good or not depends on the referring part chosen. The embedding process takes a nucleus fact where embedding would happen, a list of facts to be embedded (through applying embedding mutation), the referring parts of Arg1 and Arg2 of the nucleus fact, the available syntactic slots in both expressions (the same number of syntactic slots as is used in ILEX-TS) and all the selected rhetorical relations, and returns the numbers of good and normal embeddings according to the conditions introduced in Section 6.2.2. The remaining embeddings are treated as bad ones. This process is a part of sequence evaluation.

In order to allow associate shifting, we need to represent the relations between individual entities. In a generation system, the relation between objects are usually represented in its knowledge base as a link between them. The relations in our system have the form:

sem-rel(ancestor, descendent, reln, reln-cardinality)

where *ancestor* and *descendent* are two objects, and *reln* is the relation between them, such as whole-part, sister, etc; *reln-cardinality* gives the cardinality of the relation. For example, *sem_rel(kirkcaldy_room,ceiling,merop,1)* means that there is a whole-part relation between the room and its ceiling and the room can only have one ceiling. This is not a general way to deal with semantic relations. A more general semantic network, like KL-ONE (Brachman and Schmolze, 1985), is needed for representing the relation, which should actually hold between a supertype of *kirkcaldy_room*, like room, and ceiling in general.

When there is a semantic relation between the current entity and one mentioned in the previous discourse, a definite NP form is chosen, which represents a bridging description.

The surface realiser

Since other parts of the input are canned, the main function of the surface realiser is to generate a complex NP structure and then transfer it into natural language. The compound structure representing a noun phrase looks like:

np(np(prehead-modifiers,head-form),posthead-modifiers)

where *prehead-modifiers* are qualities to be expressed as adjectives before the NP head; *head-form* specifies the syntactic form that is chosen to realise the head; and *posthead-modifiers* are qualities to be expressed as prepositional phrases or non-restrictive relative clauses after the NP head. For example, from the following structure,

np(np('reconstructed',definite),'from the house of a prosperous burgess in Kirkcaldy')

the phrase *the reconstructed ceiling from the house of a prosperous burgess in Kirkcaldy* is generated. *definite* gives the syntactic form for realising the head and the head phrase is stored in a separate variable. The algorithm is implemented in a very simple way.

In our domain, a head noun often needs to be modified by multiple adjectives, such as *the important Scottish designer Jessie King*, so prehead adjective ordering is an indispensable task. GA-plan uses a simple ordering strategy, which is briefly described in Appendix A.3.

7.4.8 A Worked Example

This section uses a human written description to illustrate how GA-plan works. Through manually breaking down the text into individual facts and relations, we get the input to the GA system. Meanwhile, we have to make certain simplifications like transferring generic references to specific ones because GA-plan cannot handle such cases at the moment. The human text is as follows and the input to GA-plan is given in Figure 7.8 (we do not have the interestingness values of the facts, so they are set to 0).

Throne and Cover

Small portable thrones were used in the private apartments of the Imperial Palaces.

This example from the time of the Qianlong Emperor 1736-95, is made of lacquered wood with decoration in gold and red. The design on the seat is a five clawed imperial dragon in a circular medallion. On the inside of the arm pieces are small shelves on which precious possessions can be placed and studied as an aid to contemplation.

The throne cover, from the reign of Jiaquing, 1796-1820, is woven in yellow silk which is the imperial colour of the Qing Dynasty, 1644-1911. It would have covered the throne when not in use.

```
top_focus(throne).
fact(throne,be,small,fn-101,0).
fact(throne,be,portable,fn-102,0).
fact(throne,'was used in','the private apartments of the Imperial Palaces',fn-
    103,0).
fact(throne,from_date,'the time of the Qianlong Emperor 1736-95',fn-104,0).
fact(throne,material,wood,fn-105,0).
fact(wood,be,lacquered,fn-106,0).
fact(wood,has_prop,'decoration in gold and red',fn-107,0).
fact(design,location,seat,fn-109,0).
fact(design,isa,dragon,fn-110,0).
fact(dragon,be,'five clawed',fn-111,0).
fact(dragon,be,imperial,fn-112,0).
fact(dragon,be,'in a circular medallion',fn-113,0).
fact('On the inside of the arm pieces',be,'small shelves',fn-114,0).
fact('Precious possessions','can be placed in','small shelves',fn-115,0).
fact('Precious possessions','can be studied','as an aid to contemplation',fn-
    116,0).
fact(cover,from_date,'the reign of Jiaquing, 1796-1820',fn-117,0).
fact(cover,'is woven in',silk,fn-118,0).
fact(silk,be,yellow,fn-119,0).
fact(cover,'would have covered','the throne when not in use',fn-120,0).
```

Figure 7.8: The input

Figure 7.9 shows the scores of the best texts over 2000 iterations (takes about one minute). The score keeps on improving and gets stable at around 1200 iterations.

After 2000 iterations, the best text at this moment, which is scored at 86, is printed. The scoring is summarised in Table 7.1 and the RST tree of the generated text is shown in Figure 7.10. The generated text looks like:

The small portable throne from the time of the Qianlong Emperor 1736-95 is made of lacquered wood with decoration in gold and red. It was used in the private apartments of the Imperial Palaces. The cover from the reign

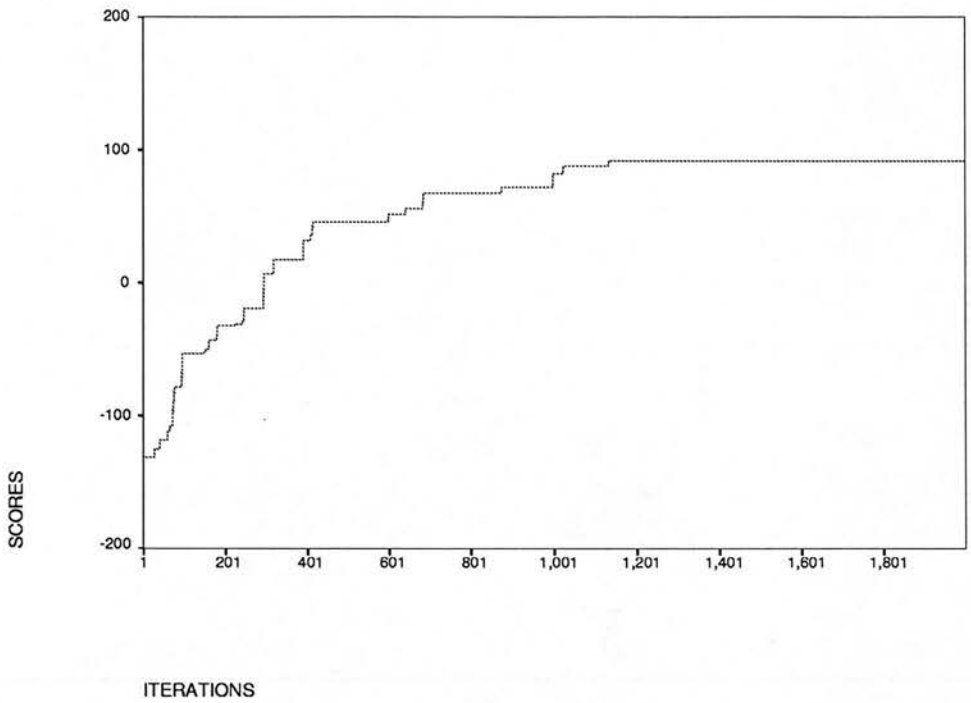


Figure 7.9: Scores of the best texts over 2000 iterations

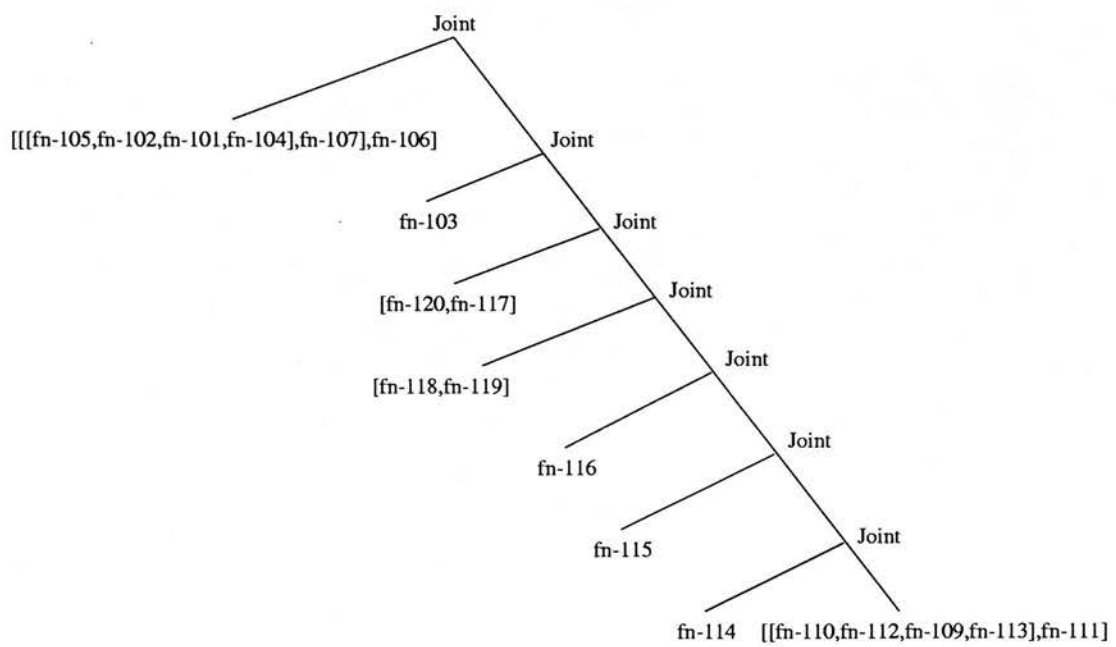


Figure 7.10: The RST tree of the generated text

Features	Scores	Explanations
joint	-20 * 7 = -140	7 JOINT relations used
embed([[[fn-105,fn-102,fn-101,fn-104],fn-107],fn-106])	10 * 5 = 50	The syntactic forms for the Arg1 and Arg2 of fn-105 are definite and indefinite respectively. All 5 embeddings are good because they can be realised as adjectives or prepositional phrases.
focus_continue(fn-105,fn-103)	20 * 1 = 20	a center continuation between the two adjacent utterances
trigger(fn-120)	16 * 1 = 16	an associate shifting from the description of the throne to that of its cover
embed([fn-120,fn-117])	10 * 1 = 10	A prepositional phrase is added to a definite description, therefore a good embedding.
focus_continue(fn-120,fn-118)	20 * 1 = 20	a center continuation
embed([fn-118,fn-119])	10 * 1 = 10	An adjective is added to an indefinite description, therefore a good embedding.
focus_continue(fn-116,fn-115)	20 * 1 = 20	a center continuation
associate(fn-114)	16 * 1 = 16	an associate shifting from the description of the throne to that of small shelves (whole-part relation)
associate(fn-110)	16 * 1 = 16	an associate shifting from the description of the throne to that of its design
intro(fn-110)	8 * 1 = 8	The first fact about the design uses the verb "is".
embed([[fn-110,fn-112,fn-109,fn-113],fn-111])	10 * 4 = 40	The syntactic forms for the Arg1 and Arg2 of fn-110 are definite and indefinite respectively. All 4 embeddings are good because they can be realised as adjectives or prepositional phrases.

Table 7.1: Summary of the scores of the best text after 2000 iterations

of Jiaquing, 1796-1820 would have covered the throne when not in use. It is woven in yellow silk.

Precious possessions can be studied as an aid to contemplation and can be placed in small shelves. On the inside of the arm pieces are small shelves. The design on the seat is an imperial five clawed dragon in a circular medallion.

7.4.9 Capturing the Rules and Preferences in GA-plan

Because a large percentage of the input to GA-plan is canned, the system does not need an ontology or an intermediate representation, except for a very simple one for NP structures. Most decisions are made according to the lexical items directly. Ignoring these differences, GA-plan uses a similar mechanism to that of ILEX-TS to conform to Rules 3.1 and 3.2. Although it does not have a set of embedding rules, the factors considered by the evaluation function to distinguish a good embedding from a bad one resemble the preconditions of embedding rules and syntactic constraints. They should be able to prevent confusing the reader about the referent indicated by the referring part and maintain the readability of an NP.

The preferences among coherence features (Heuristics 6.1 to 6.4) are captured by the relative magnitude of the numbers assigned to the features directly. This is even the case for complex feature combinations. Compared with ILEX-TS, the GA mechanism is simpler and more straightforward.

7.4.10 Summary and Discussion

This section introduces the mechanism of using a genetic algorithm for text planning, in particular the scoring of the features given in Chapter 6. Since preferred features, including good embeddings, are scored higher, they are very likely to be included in the final output. This results in reasonably readable text.

As we have mentioned, GA-plan is an experimental system and in many aspects, it is not fair to compare it with ILEX-TS, for example, on content selection and surface realisation. However, from the architecture point of view, the GA-based planner has

the following advantages over a pipeline architecture:

1. The heuristics are not hard requirements and they do not have to be incorporated in the generated text. All the properties of an RST tree are evaluated together, therefore embedding truly happens in text structuring, and the complex interactions between aggregation and planning entity-based and relation-based coherence can be modelled easily.

2. The GA architecture allows a large search space to be explored and different texts with comparable qualities to be produced.

3. In NLG, optimising individual modules does not necessarily lead to better overall performance. We believe that the GA architecture offers a mechanism for achieving good overall performance. In this architecture, various features are optimised simultaneously, so no backtracking is needed and in theory the best overall text at a given moment can always be found.

7.5 Summary

This chapter describes the implementation of the concerns of the previous chapters in two natural language generation systems: ILEX-TS and GA-plan. We present the details of how aggregation is realised in each system. One central issue of this chapter is to compare two generation architectures: a pipeline architecture, where tasks are fulfilled one by one and there is little interaction between tasks, and an architecture using a GA, where generation features are considered and optimised together so complex interactions between tasks can be captured with ease. We argue that as to modelling the complex interactions between tasks, the GA architecture is more advantageous than a pipeline structure.

However, this chapter does not show through comparing generated texts that the GA architecture is indeed better than the pipeline architecture in terms of text generation or give a clear picture of the behaviour of the GA planner. A more detailed evaluation will be given in Chapter 8.

Chapter 8

Evaluation of Preferences

Evaluation is important to demonstrating the relative superiority of NLG systems and algorithms. In Chapters 3 and 4, we evaluate the accuracy of the embedding heuristics using the annotated GNOME corpus. This chapter focuses on evaluating the coherence of multi-sentential texts generated taking into account the interactions between aggregation and text structuring. This is also one way of evaluating the GA planning architecture. We experiment with different scoring functions and automatically compare the output of GA-plan with human written texts. The evaluation shows that Heuristics 6.1 to 6.4 indeed capture some truth about the notion of a coherent text. This is further confirmed by using human subjects to assess the fluency of the texts generated by ILEX-TS and GA-plan.

8.1 What and How to Evaluate?

Evaluation is a difficult problem to NLG research. Mellish and Dale (1998) discuss the difficulties in great detail and summarise the methods that have been used. They distinguish between three types of evaluation: evaluating the system, evaluating the underlying theory and evaluating the application potential, which are however closely related. For example, evaluating a system could shed light on its underlying theory. Indeed, most NLG theories are evaluated through the systems that implement them. And the situation is that although almost all current NLG systems have some sort of evaluation, there are no generally agreed methodology or corpora that can be used to

evaluate the output of most systems and compare their performance.

Little work has been done to evaluate particular aggregations or aggregation algorithms. In particular, what is a good aggregation, what is the effect of one type of aggregation on another and is there a best order to apply aggregation rules? (Dalianis, 1996) focuses on the conciseness aspect of aggregation and calculates how much shorter a text is by using aggregation. (Shaw, 1998b) uses examples from the linguistic literature to evaluate his algorithms for segregatory coordination and ellipsis.

Since this thesis mainly studies embedding, our evaluation should focus on embedding and its effect on coherence. This section identifies the unique aspects of our theory and implementation which need further justification. It is worth pointing out that the purpose of evaluation is to provide objective judgement of a theory or system by someone other than that the author. Therefore, evaluation does not have to be a separate process after implementation, but can come in at any point of system design, as long as it is an impartial judgement. Our evaluation should have three aspects, following the distinctions made in (Mellish and Dale, 1998):

- evaluating the theory, e.g., the embedding heuristics derived from corpus analysis.
- evaluating the system, in particular GA-plan, through assessing its output. This actually evaluates the preferences among coherence features implemented in the system, which overlaps the evaluation of the theory.
- evaluating other textual effects of embedding, especially conciseness.

8.1.1 Evaluating the Theory - Embedding Heuristics

In Chapters 3, 4 and 5, we summarise several embedding rules and heuristics through corpus analysis and a psycholinguistic experiment. They are an important part of our theory about embedding and they form the basis of the embedding algorithms implemented in ILEX-TS and GA-plan.

Traditionally, a theory is tested using a few hand-crafted examples, and this lacks objectiveness and generalisation. In recent years, some effort has been made on the quantitative evaluation of a theory. An example is the work of Robin (Robin, 1994b;

Robin, 1996a), where the portability of the revision rules obtained from the analysis of corpus data is tested by applying the rules to a different corpus and a new domain. The results show that a large part of the rules are fully portable.

Unlike much other work on deriving rules from corpus analysis, which is normally based on the intuition of an individual researcher, we use independent observations besides that of the author in deriving embedding heuristics, that is, we summarise rules and heuristics from a corpus that is relatively reliably annotated and from statistical analysis of a group of independent answers to a questionnaire encoding the relevant factors. Heuristics obtained this way are more replicable in the target domain.

The evaluation of our embedding heuristics mainly concerns testing their accuracy and coverage. Since the rules cover both the selection of information to be expressed as NP modifiers and the realisation of the selected information, the evaluation of the heuristics should consist of two parts: evaluating their effectiveness in selecting information and evaluating their accuracy in realising information.

As mentioned in Section 4.3, we do not intend to address the first issue in greater detail than giving the heuristics in Section 4.3.3 because content selection is usually domain specific. Rules for one domain are not likely to be portable to a different domain.

For the second issue, we follow the same line as Robin and evaluate our embedding heuristics, in particular the decision tree for NP modifier realisation, on a part of the GNOME corpus that is not used for deriving the tree. This has been described in Section 4.3.3, where we run the *wagon_test* program on the decision tree and the annotated ICONOCLAST texts (medical information leaflets). The results show that the overall successes of the model are comparable with those of the museum domain. Since ICONOCLAST texts are more like instructions than object descriptions, this shows that the decision tree for determining NP modifier forms is portable to a completely different domain. In addition, some corpus evidence is also given to show the correctness of the heuristics concerning embedding in definite descriptions in Section 4.4.

These give us a general picture of the effectiveness of our embedding heuristics. Because the above has provided some support for the heuristics and only a small percentage of them are actually used in the implemented systems ILEX-TS and GA-plan (due to the

limited types of NP and sentence structure that the systems can produce), we will not further evaluate them through evaluating the output of the systems (the complex NP types our heuristics aim for do not appear in the output). We will also not evaluate the heuristics drawn from Chapter 5 because of the lack of semantic relations in the museum domain and publicly available annotated newspaper articles.

8.1.2 Evaluating the System - GA-plan

Mellish and Dale (1998) summarise three ways of evaluating NLG systems: accuracy evaluation, fluency/intelligibility evaluation and task evaluation. Since it is the model for aggregation and text structuring that we want to evaluate, fluency evaluation is particularly relevant.

We argue in Chapter 6 that embedding affects the coherence of a text as a whole, not just the particular phrase or sentence where it happens. This demands embedding to be evaluated in the context of evaluating the fluency or readability of the whole text. We further claim that modelling the preferences among coherence features, which take into account the interactions between embedding and other generation tasks, can result in coherent text. Since the preferences are implemented in ILEX-TS and GA-plan, evaluating our claim equals measuring the fluency of the texts produced by the two systems, which is a part of system evaluation. That is, we aim at testing the following hypotheses through evaluating the two systems.

Hypothesis 8.1 *Modelling the preferences among coherence features (Heuristics 6.1 to 6.4) in generation systems can result in coherent text.*

Hypothesis 8.2 *The way that the interactions between embedding and other generation tasks are captured in generation systems contributes to a significant difference in the coherence of the generated text.*

Hypothesis 8.2 suggests that texts generated by GA-plan are significantly more coherent than those by ILEX-TS because GA-plan is more advantageous in capturing the complex interactions between tasks.

However, measuring text coherence is difficult because using human subjects is inevitable and it is well known that there is often a lack of agreement among humans and a large number of factors might affect human judgement. Some work has been done in this respect. For example, (Lester and Porter, 1995; Lester and Porter, 1997) present the KNIGHT system for generating introductions of objects and processes in a biology domain. They use two panels for evaluation, a *writing panel* containing domain experts to write introductions for domain objects and processes and a *judging panel* containing domain experts to rate the introductions written by both human experts and KNIGHT. Judgement is given on five dimensions: overall quality and coherence, content, organization, writing style and correctness. The results show that "KNIGHT scored within "half a grade" of domain experts, and its performance exceeded that of one of the domain experts".

There is also a trend toward automatically evaluating a generated text by comparing it with an original corpus text. Bangalore et al. (2000) present several evaluation metrics and compare their performance using the corpus of Wall Street Journal articles. They judge the quality of a generated sentence in terms of the differences between it and the corresponding corpus sentence. However, their method can only measure word level differences and cannot be used for evaluating more complex text properties like embedding.

For our problem, although human judgement is indispensable, the GA-based architecture offers a possibility for a semi-automatic evaluation, which is in a way similar to the method of (Bangalore et al., 2000). The reason is that this architecture needs to evaluate the quality of a text and compare texts using a fixed set of properties in the generation process. The quality of a text is reflected by the score assigned to it and two texts can be compared through their scores. However, this depends on the evaluation function to reflect the true property of a text. Therefore, in the GA architecture evaluating the coherence of a text and validating the evaluation function address more or less the same problem.

We start with the validation of the scoring function of GA-plan and hence the preferences behind it. We use a few good texts and test if they are scored high by the evaluation function. Once the function is validated to a degree, we will be able to tell

the relative fluency of the generated texts by comparing their scores. In Section 8.2, we will describe how texts are evaluated automatically for the purpose of validation.

To further test the result of automatic evaluation and the importance of the preferences to text generation, we ask human subjects to compare texts scored differently by GA-plan as well as to compare the output of ILEX-TS and GA-plan, which allow different degrees of interaction to be captured. This is again along the same line as Robin's work (Robin and McKeown, 1996; Robin, 1996b), which evaluates the overall robustness (the percentage of text samples that can be generated without acquiring new knowledge) and scalability (the percentage of new knowledge needed to cover the generation of other samples) of the revision-based generation system STREAK by comparing it with a traditional one-pass model. However, our task is different from his in that we need to compare the coherence of the texts generated by the two systems, which is a more difficult task. In Section 8.3, we will describe how we use human judgement on the texts produced by ILEX-TS and GA-plan to test our hypotheses.

8.1.3 Evaluating Other Textual Effects

Embedding also has an effect on some overall textual properties of a text, mainly conciseness. Conciseness achieved through aggregation is normally measured through calculating how much shorter an aggregated text is compared with the non-aggregated version, e.g., (Dalianis, 1996). However, embedding does not always make a text shorter. For example, using a non-restrictive clause will not reduce sentence length although using an adjective will. Since arbitrarily long sentences are not often desirable and the increased sentence complexity is only meaningful when the text is coherent as a whole, we will not simply evaluate the textual effects of embedding in this chapter.

In the remaining sections of this chapter, we describe experiments concerning Hypotheses 8.1 and 8.2.

8.2 Justifying the Evaluation Function of GA-plan

In Section 7.4.6, we describe how to score RST trees constructed from sequences of facts and the numbers we used seem to produce reasonably good texts. However, it

is not clear how the behaviour of the planner is affected by these numbers. Will the planner work properly if the numbers change slightly?

This section describes the experiments we performed to understand the GA planning mechanism better and at the same time justify its evaluation function. We call the set of numbers for coherence features that are used in the evaluation function and satisfy all the preferences among them *a scoring system* or *a rater* as they are used to judge the quality of a text.

Different sets of numbers form different raters and their judgement of the same text might be different. We claim that what really matters in text evaluation is the preferences among features, rather than the concrete scores of features (Hypothesis 8.1). According to this hypothesis, different raters satisfying the preferences we provide would agree with each other on evaluating the coherence of texts, that is, given two texts, a better text according to one rater would tend to be scored higher by another.

In statistics, the measure of agreement is made by calculating the correlation between two or more variables. To use this statistic, we need to generate different raters and look at the distribution of the scores from the raters.

8.2.1 The Raters and Their Correlations

To generate different raters, we treat the preferences in Section 6.2.3 as constraints and feed them into a simple constraint-based program. There are no extra restrictions on the values of features except that unfavourable features are given negative values. Such features include using a JOINT relation, a top nucleus not mentioning the topic of a text, a CONJUNCT relation inside other semantic relations, an unsatisfied precondition for a relation and a bad embedding. However, this does not have to be an imperative constraint. If a feature can take a range of values, we randomly select a number in that range. We set different overall ranges (e.g., all feature values are between -100 and +150) and generate six raters this way, all of which satisfy the set of constraints. Three of them are shown in Table 8.1 (raters 1, 2 and 3), together with the one given in Section 7.4.6 (rater 0). Note that the table does not list all the features but only the major ones.

Features/Factors	Values			
	0	1 (-100..150)	2 (-50..50)	3 (-70..70)
Semantic relations				
a JOINT	-20	-46	-7	-46
a CONJUNCT or DISJUNCT	15	58	21	11
a relation other than JOINT, CONJUNCT or DISJUNCT	21	121	31	69
a CONJUNCT inside other semantic relations	-50	-51	-31	-63
a consecutive use of the same semantic relation	-50	-51	-31	-63
a precondition not satisfied	-30	-47	-11	-61
Focus moves				
a continuation	20	99	17	7
an associate shifting	16	6	15	1
a smooth shifting	14	-34	11	-3
a resumption of a previous focus	6	-37	1	-43
Embedding				
a good embedding	10	55	15	3
a normal embedding	6	51	12	0
a bad embedding	-30	-73	-17	-64
Others				
'isa' fact is the first of the text	8	32	12	7
topic not mentioned in the first sentence	-10	-7	-30	-12

Table 8.1: Four different raters satisfying the same set of constraints

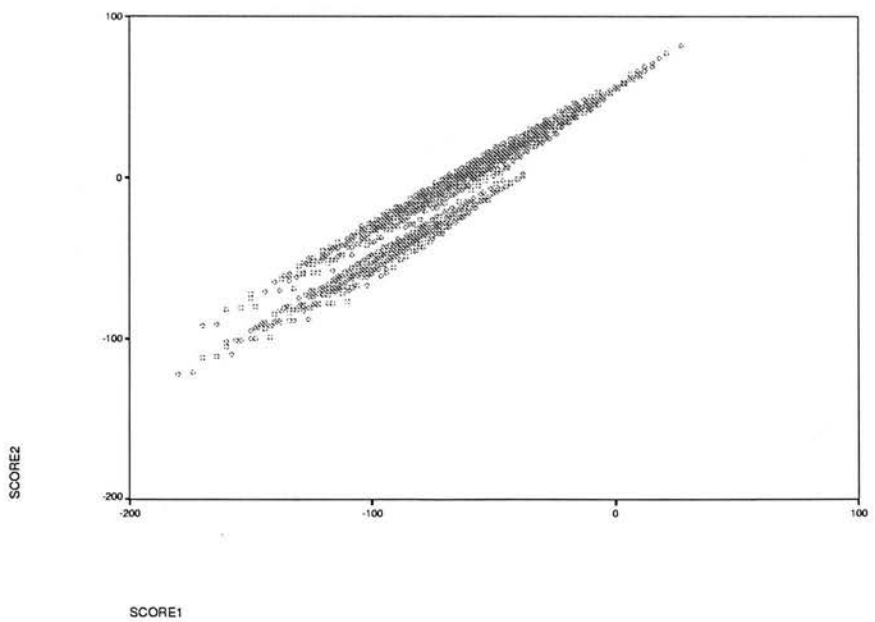


Figure 8.1: Scatterplot of scores from rater 1 and rater 2

We then broke down one human text into a number of facts and relations and chose two raters which are the number sets corresponding to raters 0 and 3 in Table 8.1. We call them rater 1 and rater 2 below.

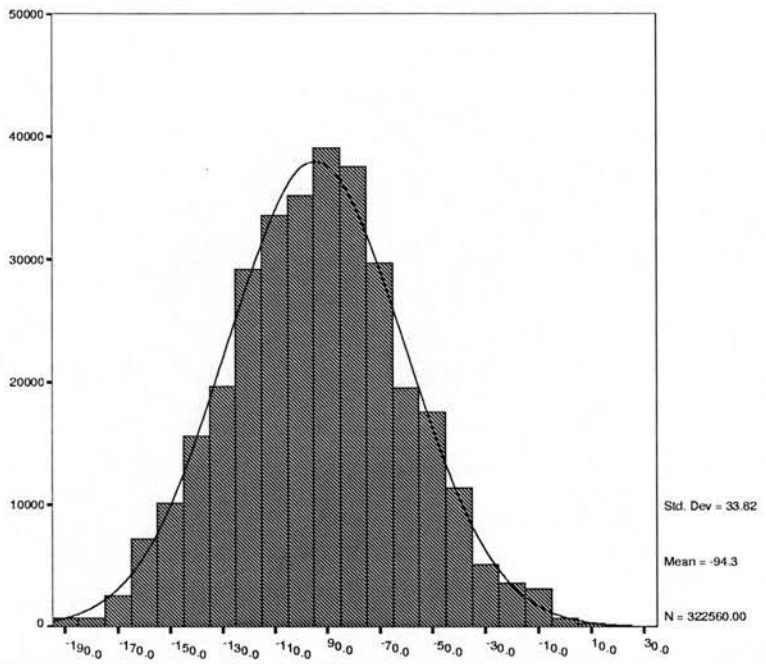
Because of the huge number of combinations, we generated all possible combinations, including embedding, of only seven facts from the text and used the two raters to score each combination. We draw a scatterplot of the scores, which not only gives a visual indication of whether there is any correlation, but also establishes whether or not the pattern of relationship is linear. The scatterplot is shown in Figure 8.1, where the axes, Score1 and Score2, represent scores rater 1 and rater 2 give to fact combinations. Each member of the text population is represented by a dot whose position is defined by the two scores given by the two raters.

Figure 8.1 shows that the points cluster tightly around a straight line with positive slope, which means that we have a strong positive linear correlation between the variables Score1 and Score2. That is, the higher the score from rater 1 for a given text of the population, the higher the score from rater 2 tends to be.

The distributions of the two scores (Figure 8.2) show that the qualities of the generated texts are of normal distributions according to both raters. The raters distinguish between good and bad texts and they classify the majority of texts as of moderate quality and only very small percentages as very good or very bad texts.

Since the means of the two distributions are quite different, the two raters assign different scores to a text. Although this is not obvious, the distribution of scores from rater 1 is slightly thinner and taller, whereas that of scores from rater 2 is a bit fatter and shorter. This can be shown from the slight difference in standard deviations, where the deviation of the latter is bigger and therefore the rater has more distinguishing power. Despite these differences, the behaviours of the two raters are indeed very similar as the two histograms are of roughly the same shape and the difference in standard deviations is not significant at all.

We are more interested in the right half of a histogram because it tells how many good texts there are and if they can be distinguished from the rest. Again the shapes of the two halves of the histograms are very similar. So the impression we get from



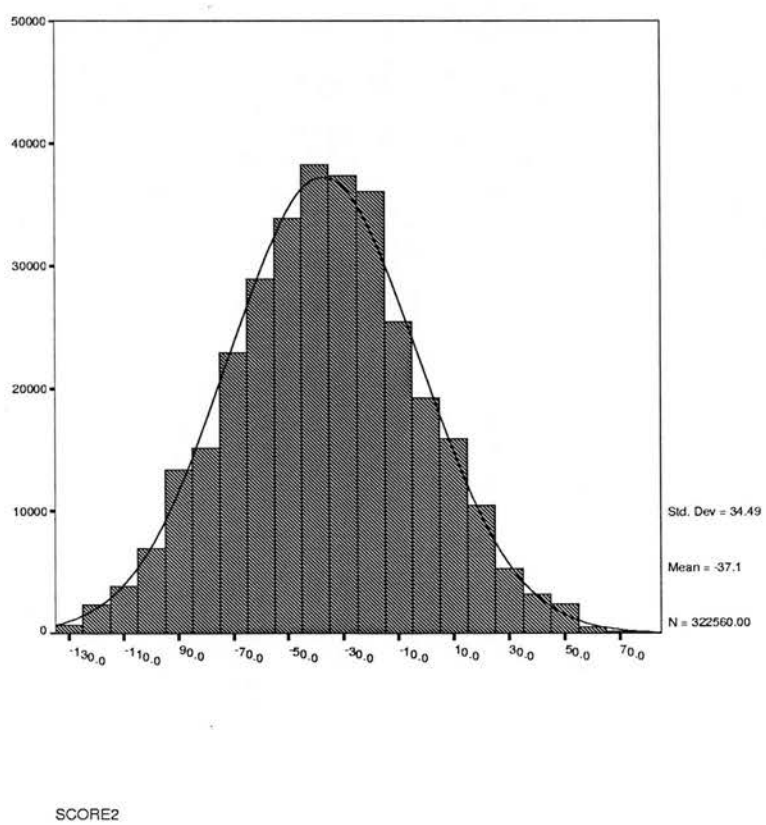


Figure 8.2: Histogram of the scores from rater 1 (top) and rater 2 (bottom)

examining the distributions of the scores from the two raters is that the raters behave very similarly in distinguishing the qualities of texts in this population.

To claim that different raters measure basically the same thing, we use the Pearson correlation coefficient to pinpoint the strength of the relationship in correlation. The results of the Pearson correlation coefficients between all pairs of the six raters are given in Table 8.2. As the correlations are high, we can claim that for this data, the scores from the raters correlate, and we have a fairly good chance to believe our hypothesis that the six raters, randomly produced in a sense, agree with each other on evaluating the given text and they measure basically the same thing.

	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	.9567	.9337	.9631	.9419	.9515
Rater2		.9435	.8819	.9280	.9185
Rater3			.8650	.8462	.9574
Rater4				.9503	.8940
Rater5					.8486

Table 8.2: Correlations between six raters

However, we admit that the experiment is limited in that we have only considered versions of one text.

8.2.2 Evaluating Human Texts

Although we have shown that the GA evaluation function captures the coherence preferences supported by linguistic theories and different raters conforming to the preferences behave in a similar way on versions of one given text, we still do not know if it indeed gives a good text a reasonably high score. The idea is that if the evaluation function ranks texts as humans do, the GA algorithm will try to find better texts through each iteration. This section describes an experiment in this respect.

Instead of using human judgement from the beginning, we try to justify the evaluation function automatically. This is possible because we have museum descriptions which are written and revised by museum experts. They could be taken as the "nearly best texts" or at least in the top 5% of texts. Since the quality of a text is reflected by its overall score, we could evaluate human written descriptions using the GA mechanism

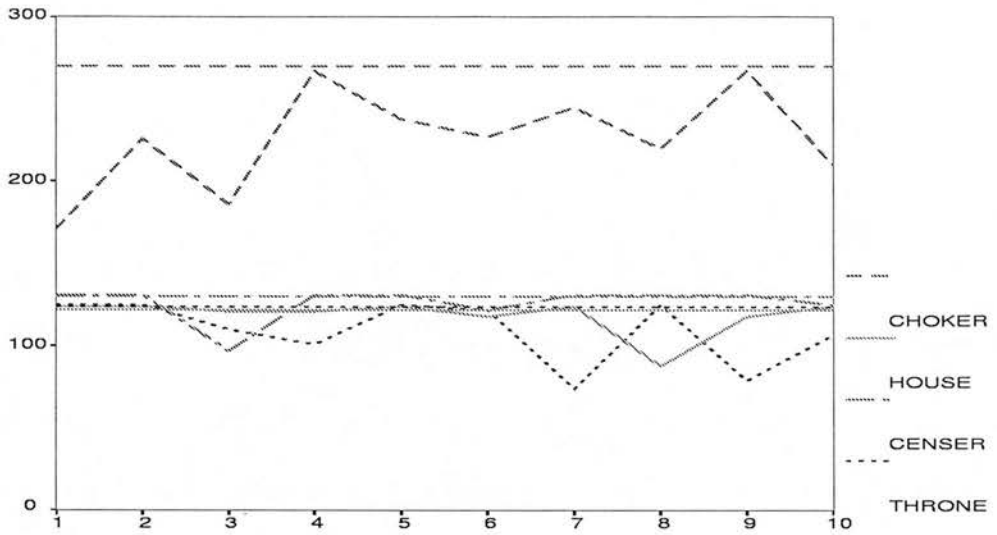


Figure 8.3: Scores for four museum descriptive texts

and compare their scores with those of the generated texts. If the former are among the highest of the scores of generated texts, we have partially validated the evaluation function and we could know how good a generated text is through comparing its score with that of the human text.

To do this, we manually broke down four human written museum descriptions into individual facts and relations and reconstructed sequences of facts with the same orderings and aggregations as in the original texts. We then used our evaluation function to score the RST trees built from these sequences. In the mean time, we ran the GA algorithm for 5000 iterations on the same input for 10 times each. The results are shown in Figure 8.3, where the four line styles correspond to the four texts. The jagged lines represent the 10 scores of the machine generated versions for each text and the straight lines represent the scores for the corresponding human texts.

All human texts are scored among the highest and machine generated texts can get scores very close to human ones sometimes. Since it is reasonable for us to believe that the scores are of normal distributions (cf. Figure 8.2), Figure 8.3 shows that the evaluation function based on our preferences can help to find good and correct combinations. So the preferences must have captured some truth about the notion of a

coherent text. The reason for a relatively bad text being generated sometimes might be that bad initial sequences are used. This could be improved by using certain heuristics to get better initial sequences. Also when the number of facts becomes larger, more iterations are needed to get readable texts. However, these are problems with the search mechanism rather than with the evaluation function.

The reason that we only experimented with four human texts is that human texts are usually much more complex than those that can be handled by current NLG systems. To enable GA-plan to compare a generated text with a human text, we have to make certain simplifications, for example, removing complex adverbial phrases, to the chosen museum descriptions in order to transfer them to a format acceptable to GA-plan. However, if too many simplifications are needed, it will not form a valid comparison between the generated and the original texts by comparing it with the simplified version. So we had to choose relatively simple descriptions and we ended up with only four such texts.

8.3 Judging Text Coherence Using Human Subjects

We mentioned in Chapter 7 that a parallel architecture (like that of GA-plan) could model the interactions within and between generation tasks better than a pipeline architecture (like that of ILEX-TS). Therefore we hypothesise that an NLG system featuring a parallel architecture generates more coherent text than that featuring a pipeline architecture (Hypothesis 8.2). In this section, we compare the two architectures by comparing the output of the two implemented systems and further validate the evaluation function of GA-plan. To do these, we need to score the coherence of the texts generated by different means.

When it comes to determining the coherence of a text, there is no better way than using human judgement. Although this method does not always give reliable results, human bias can be balanced by discovering general agreement among a relatively large number of independent observations. We use human judgement for two tasks:

1. Comparing texts scored differently by GA-plan, so that the consistency between the human and GA ranking of texts can be measured;

2. Comparing texts generated using the pipeline and parallel architectures to get an idea of their performance in text generation.

We collect texts generated by ILEX-TS and GA-plan and also texts from GA-plan with different coherence scores and put them into a questionnaire for humans to judge their coherence. The questionnaire is given in Appendix B.2.

8.3.1 The Design of the Experiment

The questionnaire contains descriptions of six jewels, a subset of the eight jewels in the ILEX domain as selected by Pearson (2000). They were chosen to get as much variety of jewel types as possible, as well as variety in their designers and styles. The reason that we use six jewels is again to obtain as large a sample as possible which can still be judged by human subjects in a relatively short period of time (less than 30 minutes). For each jewel, three texts are generated from more or less the same set of facts and relations, one by ILEX-TS and two by GA-plan. The two GA texts resemble an optimal text, which is the best result of five runs with 5000 or 8000 iterations (depending on the number of facts), and a moderate or bad text (the base line), which is the result of a run with 800 iterations. Generally the texts scored higher have more embeddings than those scored lower.

Since GA-plan is concerned mainly with the structuring of the text rather than with its realisation, the output often lacks readability due to various surface problems, for example, improper punctuation and capitalisation. In addition, we are mainly interested in the interaction between aggregation and text structuring, not grammatical phenomena unrelated to embedding. Therefore, to allow humans to focus on the targeted coherence features of the generated texts without being distracted or influenced by realisation problems, we need to hand edit the texts from GA-plan. This involves the following:

- Making proper capitalisation, punctuation, paragraph breaking and subject-verb agreement;
- Adding "usually" to propositions about generic entities, for example,

Organic style jewels usually draw on natural themes for inspiration.

- Allowing appositions involving proper names, e.g., "the British designer Jessie King". This revision is due to the simplicity of the current implementation of GA-plan, which results in something like "British Jessie King" for the first mention of a discourse entity with a proper name. In theory, there is no difficulty in generating such appositions.

- Alternating the use of pronoun (e.g., "It") and definite phrase (e.g., "The jewel") to reduce the irritation caused by the repetitive use of a pronoun, especially in the subject position.

Similar hand-editing processes are also described in (Mellish et al., 1998a; Pearson, 2000). Because the information about the jewels is very similar, we position the three descriptions of the same jewel on the same page of the questionnaire so that the subjects can compare the descriptions when they judge the coherence of each individual text. We hope that this arrangement can ease the scoring decision and also encourage the subjects to give different scores to texts. The jewels and the descriptions for each jewel are randomly ordered. The subjects should give a number between 0 and 10 to a text, representing their assessment of the coherence of the text. They are instructed to focus on how a text is arranged and ignore such factors as the use of vocabulary and repetitions between descriptions. The questionnaire is given to the subjects to be completed in their own time.

We had 10 native English speakers, all of whom work in universities, fill in the questionnaire and provide further comments about the texts if they wish. The results are analysed in the next section.

8.3.2 Results and Discussion

The null hypotheses are that there is no significant difference in coherence between the texts in each of the two groups of texts, i.e., ILEX-TS and optimal GA-plan texts, and optimal and bad GA-plan texts. Again the significance level is .05. We use the Wilcoxon Matched-Pairs Signed-Ranks Test on our data because the sample size is small and the judgement of the coherence of each text in the questionnaire is not

independent but by comparing with other two texts about the same jewel. So this is a case of repeated measures. The results of the test are summarised in Table 8.3.

Paired Variables	Cases	Z value	2-tail Sig
GA-high/GA-low	60	-5.3316	**< .0005**
ILEX-TS/GA-high	60	-1.6051	.1085

Table 8.3: The output of the Wilcoxon Matched-Pairs Signed-Ranks Test for the evaluation data

The table shows that there is a significant difference between the coherence scores given by humans to the two groups of texts scored differently by GA-plan. So there is a correlation between the human judgement of the coherence of texts and the coherence scores given by the GA evaluation function. That is, a text in the questionnaire scored higher by GA-plan tends to be favoured by humans. This means that the evaluation function of GA-plan captures similar criteria in judging text coherence to those of humans and therefore further validates our claim that the preferences behind the evaluation function capture some truth about the notion of a coherent text. The mean of the scores for the optimal GA texts is 5.7. This supports Hypothesis 8.1 to some extent, i.e., modelling the preferences among coherence features can result in texts of moderate degree of coherence.

Table 8.3 also shows that there is no significant difference between the coherence scores of the texts generated by ILEX-TS and by GA-plan. One explanation for this could be that the parallel architecture behind GA-plan can produce texts no more coherent than those generated by a pipeline architecture, and therefore capturing the interactions between tasks might not be so important to the production of coherent texts. However, if we consider that GA-plan models not much more than the interactions between aggregation and text structuring and yet is able to produce texts as coherent as ILEX texts, which are the results of several years of research and development, then capturing the interactions between tasks is indeed most important to text generation. Our current practice shows that it is too early to expect the strategy used in GA-plan to improve upon the best result of current NLG techniques, but it is certainly promising.

We analyse the comments from human subjects concerning the quality of the texts in the questionnaire to find out clues to why the parallel architecture has not performed

better. One of the most frequently complained about problem is the overuse of cue words such as "indeed" and "in that", especially when the meanings of the two propositions connected by the cue word does not authorise such a use. The reason could be that ILEX has more complex strategies for both the selection and realisation of semantic relations. For example, in content selection and structuring, ILEX assigns a weight to each relation and uses these weights to compute the best relation combination to be included in an entity chain; in content realisation, it uses an algorithm to choose suitable cue words and the appropriate nucleus/satellite positioning. In contrast, GA-plan does not have these mechanisms and uses only very simple heuristics for choosing and realising a semantic relation. For example, it penalises the consecutive use of the same relation. Some texts from GA-plan are rated low for this reason.

Another factor is that texts starting with an "isa" type proposition seem to be preferred by some subjects, whereas GA texts usually feature more flexible text structures. However by adjusting the score of this feature, it is not difficult for GA-plan to generate such texts.

It worths pointing out that ILEX-TS is based on ILEX which has been developed over several years whereas GA-plan is only an experimental system. It is likely that during the process of developing the system, the designers of ILEX put their own intuitions into the generation process so that the output has been tuned to suit the domain. Although GA-plan models some of these heuristics, this is not enough and more domain specific heuristics might have to be captured. For example, it could pose more restrictions on the choice of semantic relations and the ordering of facts and prefer more a sequence starting with an "isa" proposition. However, these are not problems with aggregation or the generation architecture.

It is also possible that the evaluation function of GA-plan only captures some coherence features but misses out some important ones. Further research is needed to establish what these features are.

In summary, our evaluation results do not support Hypothesis 8.2, that is, modeling the preferences in a better way by itself will not significantly increase the coherence of the generated text. The mean of the scores for the ILEX-TS texts are 6.0, which is

slightly higher than that of the optimal GA texts (5.7). We analyse the reasons that might contribute to this result.

Human comments also illustrate some overall properties of the texts, which explain why some subjects were able to recognise the non-human nature of the texts right away. These comments include:

- Repetitions in the use of syntactic structures and phrases in a text can be irritating. For example, our texts often have sentences like "... made in the UK, ... made in 1970, ... made from silver". Human subjects also do not like separate sentences with similar meanings, e.g., "The jewel is set with jewels. It is encrusted with gems...".

- Most of the texts have sentences that are too short, whereas human written texts often mix long sentences with short ones. Some subjects prefer to have a short sharp sentence at the start, followed by longer sentences for elaboration. They pointed out that more embedding in verbal phrases can be used to make longer sentences and reduce repetition.

These problems cause humans to rate the generated texts low. They are due to the limited ability of the generation systems rather than directly relevant to the theme of this thesis.

8.4 Comparison with a Related Work

Previously, search-based approaches have been proposed to solve subproblems of NLG. For example, (Marcu, 1997b) uses a constraint-based approach for constructing RST trees.

The work described in (Kibble, 1999; Kibble and Power, 1999; Kibble and Power, 2000) is particularly interesting to us because it discusses the relation between text planning and pronominalisation. Kibble and Power (2000) claim that text and sentence planning need to be driven in part by the goal of maintaining referential continuity and thereby facilitating pronoun resolution. This is because the effect of text and sentence

planning, including obtaining a favourable ordering of clauses and of arguments within clauses, is likely to increase opportunities for non-ambiguous pronoun use.

To maintain referential continuity in text planning, they propose the heuristics for cohesion, salience and cheapness for the task of text planning in conformity with Centering Theory. These heuristics capture the preferences for center transitions in Centering Theory but require the identification of the backward looking center (Cb) of a clause. The solution they adopt is to treat this as an optimisation problem, which gives a weight to each violation of the heuristics and tries to minimise the costs for the defects caused by all violations.

This method is implemented in the text planner of the ICONOCLAST system. The text planner takes as input a rhetorical tree whose terminal nodes are not ordered. Its task is to realise the rhetorical structure as a text structure in which propositions are ordered and if appropriate linked by cue phrases. The text planner first enumerates all acceptable text structures for a given rhetorical structure and all permissible choices of the Cb and Cp (preferred center), then it calculates the total violations of the heuristics and chooses the text structure with the smallest penalty.

This idea is similar to our GA approach in that it also uses the rules of Centering Theory for planning coherent text and it treats the maintenance of coherent center transitions as an optimisation problem in text planning. ICONOCLAST tries to minimise the cost of violations of the theory and GA-plan tries to maximise the transitions preferred by the theory, so they aim at achieving similar goals. The differences between the two approaches are mainly:

- ICONOCLAST models Centering Theory at a more refined degree than GA-plan does. The latter only implements a simplified version of the theory, for example, it does not account for the preferred center Cp. It would be an interesting future work to incorporate the planning operators of ICONOCLAST into GA-plan.

- ICONOCLAST optimises center transitions after the construction of rhetorical structures, whereas in GA-plan, the two processes are integrated and optimised at the same time.

- The disadvantage of a constraint-based approach is that it has to search the whole solution space for the global maximum/minimum, which is computationally expensive and time consuming. Such an approach can only work on small sample texts and is impractical to be used for real texts of considerable size. A GA approach has the advantage of searching through a large space and converging fast to a reasonably good solution. Although there is no guarantee that the best solution can be found, there is usually no such a requirement in NLG. Therefore, the GA approach is more practical for real world applications.

8.5 Summary

This chapter focuses on the evaluation of a major contribution of this thesis, the preferences among coherence features, which capture the interactions between different generation considerations. Since the preferences are implemented in ILEX-TS and GA-plan, the evaluation of the preferences can be fulfilled by evaluating the output of the two systems, in particular that of GA-plan.

This task somehow overlaps the task of validating the GA evaluation function which consists of two parts, automatic comparison of human texts with reconstructions of such texts using GA-plan, and human judgement of the texts scored differently by GA-plan. The results show that human subjects agree with GA-plan in judging the coherence of the sample texts. Therefore, we have confidence to believe that the evaluation function indeed captures some truth about the notion of a coherent text and capturing the interactions between generation tasks properly will lead to the production of coherent text.

We have also argued that the presence of the complex interactions between generation tasks demands a non-pipeline architecture which captures the interactions better. Using human judgement, we compare the texts generated by ILEX-TS and GA-plan, which shows that they achieve a similar degree of coherence in their output and that a well-developed NLG system using a pipeline architecture might perform slightly better than an experimental system using a non-pipeline architecture. This means that our hypothesis about the non-traditional architecture is not confirmed by the experiment.

We analyse the reasons behind this using the comments from our subjects.

This shows that at the current stage, the advantage of a parallel architecture is not its ability to improve the coherence of the generated text, but its ability to search the space of "good" texts better and allow a variety of "good" texts to be generated.

Chapter 9

Conclusions and Future Work

This last chapter summarises the important issues discussed throughout the thesis and how this thesis contributes to a better understanding of these issues. There is no doubt that the work described can be extended in many ways. Here only a few possibilities are suggested.

9.1 Main Issues Again

This thesis takes the initiative in studying how to achieve conciseness through aggregation while maintaining coherence in natural language generation. The balancing of the two conflicting considerations demands a better understanding of the interactions between generation tasks. The thesis touches several important problems of NLG, with a focus on embedding. It discusses how these problems interact with each other and what this interaction implies for the generation architecture.

In Chapter 1, we gave the central thread of the thesis, that is, the work described here is along the line of observing regularities for embedding $\rightarrow$ clarifying the interactions between embedding and other processes $\rightarrow$ extracting preferences $\rightarrow$ implementing the preferences $\rightarrow$ evaluation. This chapter re-addresses our contributions to the issues along this line, starting from the ones that we think are more important. Possible extensions to our work are also suggested.

9.1.1 Revealing the Interactions between Embedding and Document Structuring

It is mentioned in the literature that aggregation is a problem that needs clarification in almost every aspect. Some researchers suggest that there does not exist a self-contained problem called aggregation. Instead, phenomena currently classified as aggregation can be handled by a combination of existing generation processes (Wilkinson, 1995). We think that aggregation can be taken as the task of achieving conciseness through combining representations and the mess is actually caused by the complex interactions between aggregation and other generation tasks. Yet there is great difficulty in finding discussion on this topic in the literature.

Contributions

In Chapter 6, we reveal the interactions between aggregation and document structuring. We show that embedding can affect entity-based and relation-based coherence and paragraphing, and different sub-types of aggregation have an effect on one another. Therefore we believe that most aggregation phenomena, including embedding and semantic parataxis, need to be accounted in document structuring rather than just in sentence planning as is the situation in most current NLG systems.

Future Work

Besides document structuring and referring expression generation, aggregation is also closely related to lexicalisation, which has been briefly mentioned in various work on aggregation, e.g., (Dalianis, 1996). However, no detailed discussion on how the two tasks interact can be found. It would be very revealing to discuss the issue and what it suggests for natural language generation.

9.1.2 Modelling the Interactions between Generation Tasks

Although preferences among some coherence features have been proposed in the literature of discourse analysis, e.g., those for local coherence as described in (Grosz et al., 1995), the features that are considered are usually for a single phenomenon.

Contributions

In Chapter 6, we present a novel way of modelling the interactions between generation tasks, i.e., to capture them as preferences among coherence features. These preferences are not restricted to a specific phenomenon, but cross phenomena. This provides a unified way of modelling interactions at all levels, both within and between tasks. More preferences can be easily incorporated into this framework and conflicts between preferences can be detected.

We describe the implementation of the preferences in two generation systems in Chapter 7. We compare the ways the preferences are captured and list problems with the pipeline architecture. The implementations show that the non-pipeline architecture (GA-plan) models the preferences more naturally, whereas the pipeline architecture (ILEX-TS) has to use more complex algorithms to realise the preferences.

Future Work

In Chapter 3, we demonstrate the interactions between the referring and embedding processes, but we have not included them into the set of preferences. Work in this respect will enhance the preference set significantly.

Although our preferences consider center transitions in document structuring (Section 6.2), the problem of how to model different types of transition is left open and our implementations realise Centering Theory in a very simple way. We introduced in Section 8.4 that (Kibble and Power, 2000) uses a constraint-based approach to optimise the preferences of Centering Theory in text planning. It will benefit our implementations to incorporate their planning operators.

We have also mentioned that the set of preferences is not intended to be a complete set. It is possible that we have missed out some important features. Therefore, incorporating preferences from others' work into this framework is useful. In addition, it would be interesting to see if these preferences can be applied to other domains, where more complex rhetorical phenomena exist. We would also expect more domain specific preferences to be added when it comes to generation problems in a specific domain.

9.1.3 Generating Complex Referring Expressions

The observation that human produced NPs often fulfill multiple functions poses the question of how to automatically generate NPs with comparable capability. The interaction between aggregation and referring expression generation is again the central issue that needs to be clarified.

Contributions

We divide the components of a referring expression into a referring and a non-referring part. We identify three functions that an NP can serve in human written descriptions: referring to a discourse entity, supporting the situation described in the main proposition containing the NP and providing additional information about the referent. Our focus is on the construction of the non-referring part, which serves one of the last two functions.

To generate a non-referring part, we discuss in Chapter 3 the bilateral relation between the referring and non-referring part of an RE, which causes the complex interaction between the referring process and embedding. We propose Rules 3.1 to 3.3 as general restrictions on embedding, which require that an embedding should not cause confusion, reduce readability of the text as a whole or change the property of the referent. In particular, we discuss how to conform to the first rule, including such issues as safe embedding in bridging descriptions and other types of definite descriptions, which have not been addressed by previous work.

In Chapter 5, we take a first step toward finding out the factors that decide the use of NP modifiers to support the meaning of the main proposition. Our experiment uses statistical analysis and gives reliable evidence that when certain conditions (e.g., relation, inferrability) are satisfied in the input, using NP subordination can be a good way of expressing a semantic relation, which is normally verbalised in NLG systems through separate clauses connected by a cue phrase.

The rules and heuristics we present enable an NLG system to generate complex referring expressions that are capable of serving not only their primary function of referring to discourse entities but also other secondary functions without disrupting the main

function.

Future Work

In our work, we assume that a modifier can only serve one function or it always has a primary function. However, during the annotation of the GNOME corpus, we have found that an NP modifier can sometimes be assigned multiple *PRAGM* values. If an NP generation algorithm can choose among multiple possibilities those properties that serve more than one function, the generated NP will be more efficient and informative.

In addition, the algorithm we give for capturing the interactions between referring and embedding is only a simplified solution. A better algorithm is desirable, for example, using a constraint-based approach.

9.1.4 Deriving Embedding Heuristics

Aggregation often makes use of explicit or implicit rules to combine semantic representations. In Chapter 2, we summarise three ways of deriving rules for aggregation: using linguistic observations, psycholinguistic evidence and corpus analysis. Corpus analysis is the most frequently used method in recent research on aggregation. However, the analysis is often performed by an individual researcher and therefore the rules are based on his/her intuition. Such rules lack reliability and it might not be possible to replicate the intuition.

Contributions

We use all three means to derive embedding heuristics and emphases the sound empirical basis of the heuristics. In Chapter 4, we describe corpus analysis based on corpus annotation by multiple independent observers. Through measuring the agreement on the *PRAGM* feature, we show that the distinctions concerning modifier usage can be detected by naive human annotators to some extent. Since the corpus is annotated after reasonable degrees of agreement are achieved, the corpus can be useful for other researchers to work on similar problems. The statistical model trained on this corpus gives relatively reliable evidence for modifier realisation. In addition, we describe

tendencies concerning modifier content selection in the corpus.

Chapter 4 also describes the evaluation of some of the embedding heuristics using the annotated GNOME corpus. Our approach allows the statistical model for embedding to be tested easily in a different domain and provides reliable results as long as the domain corpus is reliably annotated with the same set of features.

The psycholinguistic experiment described in Chapter 5 also gives reliable evidence on the usage of NP modifiers to support the situation in the main proposition. The embedding heuristics obtained through all these means are more reliable and are replicable. We believe that similar approaches can be very useful for other generation oriented problems.

Future Work

Generally to make conclusive statements about a problem, the agreement in terms of Kappa statistics needs to be higher than .8. Using the current annotation scheme it has not been able to reach this level. Some revision of the scheme is needed, for example, allowing multiple values for the *PRAGM* feature. In addition, the accuracy of the statistical model in predicting modifier realisation forms may be increased by training on a larger corpus, which requires more texts to be annotated.

Many aspects of the psycholinguistic experiment can also be extended, for example, using more cue phrases to cover a wider range of each semantic relation, analysing more semantic relations that are expressed using NP subordination in a specific domain, and measuring the similarity between an NR construction and a hypotactic construction that does not use a cue phrase.

An approach similar to that of (Rambow et al., 2001), which compares several sentence planners by scoring their surface output, can be used to evaluate the embedding heuristics and reveal their effectiveness.

9.1.5 Evaluating Text Coherence

Previous evaluation of aggregation often focuses on conciseness achieved through aggregation. How it affects the coherence of a text as a whole has not been studied. The main reason lies in the difficulty in judging coherence.

The evaluation strategy we adopted is to bring evaluation into system design as early as possible rather than just treating it as the final stage of system development.

Contributions

The justification of the evaluation function of GA-plan sheds light on possibilities for semi-automatic evaluation of text coherence, which has not been touched on by previous work. Using human written descriptions, we are able to compare them with machine generated texts through the same evaluation mechanism. This partially validates the evaluation function and therefore the preferences among coherence features.

In addition to automatic evaluation, we also use human subjects to judge the coherence of texts generated by both ILEX-TS and GA-plan. This not only further justifies the set of preferences and the evaluation function of GA-plan, but also compares the two generation architectures through comparing their output. The results make us believe that the preferences capture some essential text coherence features and the performance of a non-traditional generation architecture which models these preferences matches that of a pipeline architecture at the current stage.

Future Work

We faced many difficulties during the evaluation process. We used only four texts for automatic evaluation, which is far from enough. The reason for such a small quantity is that human texts are usually too complex for GA-plan to handle. This problem can be relieved by improving the capability of GA-plan to make it a real generation system which can handle more complex generation phenomena.

In Chapter 8, we generated six raters and measured their correlation on judging only one text. This is not enough for making serious statement. It would be desirable to

generate more raters to evaluate more texts and see if they correlate. Conflicts among raters might show up in this process and it is valuable to analyse what causes the conflicts.

9.1.6 A Better Understanding of Aggregation

This thesis starts with aggregation and gradually moves to more general issues like planning text coherence. The intention is not just to study aggregation as a specific generation phenomenon, but also to see what it has to say for the generation problem as a whole.

Contributions

The thesis contributes to a better understanding of aggregation by clarifying some embedding phenomena and how embedding interacts with other generation tasks. We address the problems discussed in Section 1.1.3 from the point of view of embedding and give our answers to them. The rules and heuristics given in this thesis are summarised in Appendix A.1.

Future Work

For aggregation, there are some questions that current NLG theories cannot give a satisfactory answer to. For example, how not to change the semantics when aggregating, and how aggregation is related to stylistic considerations. The answers to these questions rely on developments in research on both aggregation and the more difficult topics of lexical semantics and stylistics.

9.2 Concluding Remarks

In this thesis, we present a different perspective to aggregation, which represents our answers to the questions discussed in Section 1.1.3. As to the questions raised by Reape and Mellish (1999), we mainly answer the "when is it done" and "what is it done to/on" questions. That is, aggregation should be taken account of while planning

entity-based and relation-based coherence, and it needs abstract syntactic information in order to make a decision. The more general question of what is the relationship between generation processes subsumes the "in what order are its subparts done" question, which again emphasises the interaction between generation tasks and motivates an architecture which allows more interactions to be captured.

We believe that modelling the interactions among coherence features is the key to the generation of a coherent text. The discussion in this thesis contributes to the understanding of not only the aggregation phenomenon, but also text coherence and generation architectures. These discussions will also be very helpful in studying interactions among other generation tasks.

References

Evelyn Abberton. 1977. Nominal group premodification structures. In Wolf-Dietrich Bald and Robert Ilson, editors, *Studies in English Usage: The Resources of a Present-Day English Corpus for Linguistic Analysis*, pages 29–72. Frankfurt/M : P. Lang.

Douglas Appelt. 1982. Planning natural-language utterances. In *Proceedings of the 2nd National Conference on Artificial Intelligence (AAAI)*, pages 59–62.

Douglas Appelt. 1985a. Planning English referring expressions. *Artificial Intelligence*, 26:1–33.

Douglas Appelt. 1985b. Some pragmatic issues in the planning of definite and indefinite referring expressions. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 198–203.

Douglas Appelt. 1987. Reference and pragmatic identification. In *Theoretical Issues in Natural Language Processing (TINLAP-3)*, pages 149–153, New Mexico State University.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG)*, pages 1–8, Mitzpe Ramon, Israel.

John Bateman, Robert Kasper, Johanna Moore, and Richard Whitney. 1990. A general organization of knowledge for natural language processing: the PENMAN upper model. Technical report, USC/Information Sciences Institute, Marina del Rey, California.

John Bateman, Renate Henschel, and Fabio Rinaldi. 1995. The generalized upper model 2.0: Documentation. Technical report, GMD/Institut fuer Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

John Bateman, Thomas Kamps, Jorg Kleinz, and Klaus Reichenberger. 1998. Communicative goal-driven NL generation and data-driven graphics generation: an architectural synthesis for multimedia page generation. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 8–17, Ontario, Canada.

John Bateman. 1995. KPML: The KOMET-Penman (multilingual) development environment: Support for multilingual linguistic resource development and sentence generation. Technical report, Institut fuer Integrierte Publikations- und Informationssysteme, GMD, Darmstadt.

Ronald Brachman and James Schmolze. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216.

Leo Breiman, J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Belmont, Calif.: Wadsworth International.

Susan Brennan, Marilyn Walker Friedman, and Carl Pollard. 1987. A centering apporach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford, CA.

Jean Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Hua Cheng and Chris Mellish. 2000. Capturing the interaction between aggregation and text planning in two generation systems. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 186–193, Mitzpe Ramon, Israel.

Hua Cheng. 1998. Embedding new information into referring expressions. In *Proceedings of COLING-ACL'98*, pages 1478–1480, Montreal, Canada.

Hua Cheng. 1999. The annotation scheme manual for NP heads and modifiers. Technical report, Division of Informatics, the University of Edinburgh.

H.H. Clark. 1977. Bridging. In Philip Johnson-Laird and Peter Wason, editors, *Thinking: Readings in Cognitive Science*, pages 9–27. Cambridge: Cambridge University Press.

Jennifer Coates. 1977. A corpus study of modifiers in sequence. In Wolf-Dietrich Bald and Robert Ilson, editors, *Studies in English Usage: The Resources of a Present-Day English Corpus for Linguistic Analysis*, pages 9–27. Frankfurt/M : P. Lang.

E. Coleman. 1962. Improving comprehensibility by shortening sentences. *Journal of Applied Psychology*, 46(2):131–134.

Robert Dale and Nicholas Haddock. 1991a. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

Robert Dale and Nicholas Haddock. 1991b. Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 161–166, Berlin.

Robert Dale and Ehud Reiter. 1994. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Robert Dale. 1986. The pronominalization decision in language generation. DAI Research Paper 276, Department of Artificial Intelligence, the University of Edinburgh.

Robert Dale. 1987. The generation of subsequent referring expressions in structured discourses. Technical Report EUCCS/RP- ; 20, Centre for Cognitive Science, the University of Edinburgh.

Robert Dale. 1990. Generating recipes: an overview of Epicure. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*, pages 229–255. Academic Press Limited.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* The MIT Press.

Hercules Dalianis and Eduard Hovy. 1996. Aggregation in natural language generation. In G. Adorni and Michael Zock, editors, *Trends in Natural Language Generation: an Artificial Intelligence Perspective, EWNLG'93*, pages 88–105. Springer Verlag.

Hercules Dalianis. 1995. Aggregation, formal specification and natural language generation. In *Proceedings of the 1st International Workshop on the Applications of Natural Language to Data Bases*, pages 135–149, Versailles, France.

Hercules Dalianis. 1996. *Concise Natural Language Generation from Formal Specifications.* Ph.D. thesis, Department of Computer and Systems Sciences, The Royal Institute of Technology and Stockholm University, Sweden.

Hercules Dalianis. 1997a. Natural language aggregation and clarification using cue words. In *Proceedings of the 6th European Workshop on Natural Language Generation*, pages 6–16, Duisburg, Germany.

Hercules Dalianis. 1997b. On lexical aggregation and ordering. In *Proceedings of the 6th European Workshop on Natural Language Generation*, pages 17–27, Duisburg, Germany.

L. Davis. 1991. *Handbook of Genetic Algorithm.* Van Nostrand Reinhold.

K. De Jong. 1975. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems.* Ph.D. thesis, University of Michigan.

Reunion des Musees Nationaux, editor. 1993. *Louvre: The Collections.* Editions de la Reunion des musees nationaux, Paris.

George Dillon, 1981. *Constructing Texts: Elements of a Theory of Composition Style*, chapter 6. Bloomington: Indiana University Press.

Chrysanne DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–499.

K. Donnellan. 1977. Reference and definite descriptions. In Stephen Schwartz, editor, *Naming, Necessity, and Natural Kinds*, pages 42–65. Cornell University Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press.

Armin Fiedler and Xiaorong Huang. 1995. Aggregation in the generation of argumentative texts. In *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, The Netherlands.

Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7:395–433.

T.R. Girill. 1991. Information chunking as an interface design issue for full-text databases. In Martin Dillon, editor, *Interfaces for Information Retrieval and Online Systems*, pages 149–158. Greenwood Press, New York.

David Goldberg. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley Publishing Company, Inc.

Peter Gordon, Barbara Grosz, and Laura Gillion. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Nancy Green, Giuseppe Carenini, and Johanna Moore. 1998. A principled representation of attributive descriptions for generating integrated text and information graphics presentations. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 18–27, Niagara-on-the-Lake, Canada.

Barbara Grosz and Candace Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. Technical Note 292, SRI International, University of Pennsylvania.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Barbara Grosz. 1977. The representation and use of focus in dialogue understanding. Technical report 151, SRI International, University of Pennsylvania.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English.* Longman, London.

M.A.K. Halliday. 1985. *An Introduction to Functional Grammar.* Edward Arnold (Publishers) Ltd., London.

Evelyn Hatch and Anne Lazaraton. 1991. *The Research Manual: Design and Statistics for Applied Linguistics.* Newbury House Publishers.

John Hawkins. 1978. *Definiteness and Indefiniteness: a Study in Reference and Grammaticality Prediction.* Croom Helm, London.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Saarbrucken, Germany.

Helmut Horacek. 1990. The architecture of a generation component in a complete natural language dialog system. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 193–227. Academic Press, New York.

Helmut Horacek. 1992. An integrated view of text planning. In R. Dale, E. Hovy, and D. Rosner, editors, *Aspects of Automated Natural Language Generation*, pages 29–44. Springer-Verlag.

Helmut Horacek. 1995. More on generating referring expressions. In K. De Smedt, C. Mellish, and H. Novak, editors, *Proceedings of the 5th European Workshop on Natural Language Generation*, pages 43–58, Leiden, The Netherlands.

Helmut Horacek. 1997. An integrative algorithm for generating referential descriptions. In *Proceedings of the 6th European Workshop on Natural Language Generation*, pages 38–49, Duisburg, Germany.

Eduard Hovy. 1988. Planning coherent multisentential text. Technical Report ISI/RS-88-208, Information Sciences Institute, University of Southern California.

Eduard Hovy. 1989. Approaches to the planning of coherent text. Technical Report ISI/RS-89-245, Information Sciences Institute, University of Southern California.

Eduard Hovy. 1990. Unresolved issues in paragraph planning. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 17–45. Academic Press.

Eduard Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence: Special Issue on Natural Language Processing*, 63(1-2):341–385.

Xiaorong Huang and Armin Fiedler. 1996. Paraphrasing and aggregating argumentative text using text structure. In *Proceedings of the 8th International Workshop on Natural Language Generation*, Herstmonceux Castle, UK.

Carla Huls, Edwin Bos, and Wim Claassen. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79.

Pamela Jordan, Bonnie Dorr, and John Benoit. 1993. A first-pass approach for evaluating machine translation systems. *Machine Translation*, 8(1-2):49–58.

Aravind Joshi and X. Schabes. 1992. Tree adjoining grammars and lexicalized grammars. In M. Nivat and A. Podelski, editors, *Tree Automata and Language*. Elsevier.

M. A. Just and P. A. Carpenter. 1987. *The Psychology of Reading and Language Comprehension*. Allyn & Bacon, Boston.

Robert Kasper and Richard Whitney. 1989. SPL: A sentence plan language for text generation. Technical report, Information Sciences Institute, University of Southern California.

Gerard Kempen. 1991. Conjunction reduction and gapping in clause-level coordination: An inheritance-based approach. *Computational Intelligence*, 7(4):357–360.

Rodger Kibble and Richard Power. 1999. Using centering theory to plan coherent texts. In *Proceedings of the 12th Amsterdam Colloquium*.

Rodger Kibble and Richard Power. 2000. An integrated framework for text planning and pronominalisation. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 77–84, Mitzpe Ramon, Israel.

Rodger Kibble. 1999. Cb or not cb? centering theory applied to nlg. In *Proceedings of the ACL Workshop on Discourse and Reference Structure*.

Richard Kittredge, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314.

Alistair Knott and Mick O'Donnell. 1998. WAG/ILEX user manual. Technical report, Department of Artificial Intelligence and Human Communication Research Centre, the University of Edinburgh. URL: cirrus.dai.ed.ac.uk:8000/ilex/Manual/index.html.

Alistair Knott, Jon Oberlander, Mick O'Donnell, and Chris Mellish. in press. Beyond elaboration: the interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*. Benjamins, Amsterdam.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, the University of Edinburgh, Edinburgh.

Amichai Kronfeld. 1990. *Reference and Computation*. Studies in Natural Language Processing. Cambridge University Press.

James Lester and Bruce Porter. 1995. The KNIGHT experiments: Empirically evaluating an explanation generation system. In *Working Notes of the AAAI Symposium on Empirical Methods in Discourse Interpertation and Generation*, AAAI Spring Symposium Series, pages 74–80, Stanford University.

James Lester and Bruce Porter. 1997. Developing and empirically evaluating robust explanation generation: the KNIGHT experiments. *Computational Linguistics*, 23(1):65–101.

W. Levelt. 1989. *Speaking - From Intention to Articulation*. MIT Press.

Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Sebastian Loebner. 1987. Definites. *Journal of Semantics*, 4:279–306.

William Mann and Christian Matthiessen. 1985. Demonstration of the NIGEL text generation computer program. In James Benson and William Greaves, editors, *Systemic Perspectives on Discourse*, pages 50–83. Norwood: Ablex.

William Mann and James Moore. 1980. Computer as author – results and prospects. Technical Report ISI/RR-79-82, Information Sciences Institute, University of Southern California.

William Mann and James Moore. 1981. Computer generation of multiparagraph English text. *American Journal of Computational Linguistics*, 7(1):17–29.

William Mann and Sandra Thompson. 1987a. Rhetorical structure theory: A framework for the analysis of text. Technical Report ISI/RR-87-185, Information Sciences Institute, University of Southern California.

William Mann and Sandra Thompson. 1987b. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RR-87-190, Information Sciences Institute, University of Southern California.

William Mann and Sandra Thompson. 1987c. Rhetorical structure theory: Description and construction of text structures. Technical Report ISI/RS-86-174, Information Sciences Institute, University of Southern California.

William Mann. 1983. An overview of the Penman text generation system. Technical Report ISI/RR-83-114, USC/Information Sciences Institute, Marina del Rey, CA.

Daniel Marcu. 1997a. From local to global coherence: A bottom-up approach to text planning. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 629–635, Providence, Rhode Island.

Daniel Marcu. 1997b. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Christian Matthiessen and Sandra Thompson. 1987. The structure of discourse and subordination. Technical Report ISI/RS-87-183, Information Sciences Institute, University of Southern California.

Kathleen McCoy and Jeannette Cheng. 1991. Focus of attention: Constraining what can be said next. In C. Paris, W. Swartout, and W. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 103–124. Kluwer Academic Publishers.

David McDonald and F. Busa. 1994. On the creative use of language: the form of lexical resources. In *Proceedings of the 7th International Workshop on Natural Language Generation*, Kennebunkport, Maine, USA.

Kathleen McKeown, Karen Kukich, and James Shaw. 1994. Practical issues in automatic document generation. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 7–14, Stuttgart, Germany.

Kathleen McKeown, Shimei Pan, James Shaw, Desmond Jordan, and Barry Allen. 1997. Language generation for multimedia healthcare briefings. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 277–282, Washington D.C.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12:349–373.

Chris Mellish and Roger Evans. 1989. Natural language generation from plans. *Computational Linguistics*, 15(4):233–249.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998a. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 98–107, Ontario, Canada.

Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. 1998b. An architecture for opportunistic text generation. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 28–37, Ontario, Canada.

Marie Meteer. 1991. Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304.

Marie Meteer. 1992. *Expressibility and the Problem of Efficient Text Planning.* Communication in Artificial Intelligence. Pinter Publishers Limited, London.

Charles Meyer. 1992. *Apposition in Contemporary English.* Cambridge University Press, Cambridge.

Melanie Mitchell. 1996. *An Introduction to Genetic Algorithms.* The MIT Press.

Johanna Moore and Cecile Paris. 1994. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694.

Johanna Moore and Martha Pollack. 1992. A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Megan Moser and Johanna Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Jon Oberlander, Mick O'Donnell, Alistair Knott, and Chris Mellish. 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.

Mick O'Donnell, Hua Cheng, and Janet Hitzeman. 1998. Integrating referring and informing in NP planning. In *Proceedings of COLING-ACL'98 Workshop on the Computational Treatment of Nominals*, pages 46–56, Montreal, Canada.

Michael O'Donnell. 1994. *Sentence Analysis and Generation: a Systemic Perspective.* Ph.D. thesis, Department of Linguistics, University of Sydney.

Franck Panaget. 1994a. The micro-planning component of a generation system. Technical Report NT/LAA/TSS/525, France Telecom, CNET.

Franck Panaget. 1994b. Using a textual representational level component in the context of discourse or dialogue generation. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 127–136, Kennebunkport, Maine, USA.

Franck Panaget. 1997. Micro-planning: a unified representation of lexical and grammatical resources. In *Proceedings of the 6th European Workshop on Natural Language Generation*, pages 97–106, Duisburg, Germany.

Mark Pearson. 2000. Determining a measure of textual coherence. BSc Dissertation, Department of Psychology, the University of Edinburgh.

Massimo Poesio. 2000a. Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, pages 211–218, Athens, Greece.

Massimo Poesio. 2000b. The GNOME annotation scheme manual. Technical Report URL: www.cogsci.ed.ac.uk/poesio/GNOME/anno_manual_4.html, Division of Informatics, the University of Edinburgh.

C. Pollard and I. Sag. 1994. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago.

Ellen Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In William Mann and Sandra Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text*, pages 295–325. John Benjamins Publishing Company.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Grammar of Contemporary English.* Longman Group Ltd.

Owen Rambow, Monica Rogati, and Marilyn Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 434–441, Toulouse, France.

Victor Raskin and Irwin Weiser, 1987. *Language and Writing: Applications of Linguistics to Rhetoric and Composition*, chapter 3. ABLEX Publishing Corporation, Norwood, New Jersey.

Michael Reape and Chris Mellish. 1999. Just what is aggregation anyway? In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 20–29, Toulouse, France.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Ehud Reiter and Chris Mellish. 1992. Using classification to generate text. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 265–271, Newark, Deleware.

Ehud Reiter. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 163–170, Kennebunkport, Maine, USA.

Ehud Reiter. 1995. NLG vs. templates. In *Proceedings of the 5th European Workshop on Natural Language Generation*, Leiden, The Netherlands.

Jacques Robin and Eloi Favero. 2000. Content aggregation in natural language hypertext summarization of OLAP and data mining. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 124–132, Mitzpe Ramon, Israel.

Jacques Robin and Kathy McKeown. 1993. Corpus analysis for revision-based generation of complex sentences. In *Proceedings of the 11th National Conference on Artificial Intelligence*, pages 365–372, Washington DC, USA.

Jacques Robin and Kathy McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, 85(1-2):135–179.

Jacques Robin. 1993. A revision-based generation architecture for reporting facts in their historical context. In Helmut Horacek and Michael Zock, editors, *New Concepts in Natural Language Generation: Planning, Realization and Systems*, pages 238–268. Frances Pinter, London and New York.

Jacques Robin. 1994a. Automatic generation and revision of natural language summaries providing historical background. In *Proceedings of the 11th Brazilian Symposium of Artificial Intelligence*, Fortaleza, CE, Brazil.

Jacques Robin. 1994b. *Revision-based Generation of Natural Language Summaries Providing Historical Background: Corpus-based Analysis, Design, Implementation and Evaluation.* Ph.D. thesis, Computer Science Department, Columbia University.

Jacques Robin. 1996a. Evaluating the portability of revision rules for incremental summary generation. In *Proceedings of the 34th Annual Meeting of the Association for computational Linguistics*, pages 205–214, Santa Cruz, CA.

Jacques Robin. 1996b. Evaluating the robustness and scalability of revision-based natural language generation. In *Proceedings of the 13th Brazilian Symposium on Artificial Intelligence*, Curitiba, PN, Brazil.

J. David Schaffer, Richard Caruana, Larry Eshelman, and Rajarshi Das. 1989. A study of control parameters affecting online performance of genetic algorithms for function optimization. In *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 51–60, George Mason University.

Roger Schank. 1977. Rules and topics in conversation. *Cognitive Science*, 1(1):421–441.

Donia Scott and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press.

James Shaw and Vasileios Hatzivassiloglou. 1999. Ordering among premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 135–143, College Park, Maryland.

James Shaw and Kathleen McKeown. 1997. An architecture for aggregation in text generation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence, Poster Session*, Nagoya, Japan.

James Shaw and Kathleen McKeown. 2000. Generating referring quantified expressions. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 100–107, Mitzpe Ramon, Israel.

James Shaw. 1995. Conciseness through aggregation in text generation. In *Proceedings of the 33rd Annual Meeting of Association for Computational Linguistics*, pages 329–331, MIT, Cambridge, Massachussetts.

James Shaw. 1998a. Clause aggregation using linguistic knowledge. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 138–147, Niagara-on-the-Lake, Canada.

James Shaw. 1998b. Segregatory coordination and ellipsis in text generation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1220–1226, Montreal, Canada.

Sidney Siegel and John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences.* London:McGraw-Hill.

Mark Steedman. 1996. *Surface Structure and Interpretation.* MIT Press, Cambridge, Massachusetts.

William Strunk and E.B. White. 1979. *The Elements of Style.* MacMillan Publishing Co., Inc.

Paul Taylor, Richard Caley, Alan Black, and Simon King. 1999. Edinburgh speech tools library system documentation edition 1.2. Technical Report URL: www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/, Centre for Speech Technology Research, the University of Edinburgh.

Renata Vieira. 1997. *Definite Description Processing in Unrestricted Text.* Ph.D. thesis, Centre for Cognitive Science, the University of Edinburgh.

John Wilkinson. 1995. Aggregation in natural language generation: Another look. Technical report, Computer Science Department, University of Waterloo.

Ching-Long Yeh and Chris Mellish. 1997. An empirical study on the generation of anaphora in Chinese. *Computational Linguistics*, 23(1):169–190.

Wlodek Zadrozny and Karen Jensen. 1991. Semantics of paragraphs. *Computational Linguistics*, 17(2):171–209.

Appendix A

Rules and Heuristics

A.1 Summary of Rules and Heuristics

Below lists all the rules and heuristics given in this thesis:

Rule 3.1 : The non-referring part should not confuse the reader about the referent indicated by the referring part.

Rule 3.2 : The non-referring part should not reduce the readability of the text. This includes three aspects:

- The referring expression should not be too complex to read.
- The non-referring part should not reduce the entity-based coherence of a discourse.
- The non-referring part should not reduce the relation-based coherence of a discourse.

Rule 3.3 : The non-referring part should not change the properties of the referent.

Heuristic 4.1 Bridging Heuristic: When the referent (R) has a trigger (T) which is mentioned in the previous discourse, and the cardinality of the association relation between the upper-model concepts immediately subsuming T and R is one, embedding will not have a side effect.

Heuristic 4.2 Salience Heuristic: When the referent is the most salient object in its context among objects of the same type, embedding would not have a side effect.

Heuristic 5.1 : An NR construction is an acceptable realisation for the specific causal or temporal relation when

- The causal relation holds between two facts and the inferrability of the relation is strong, in which case satellite subordination should be used and the cause is preferred to be given first; or
- The temporal relation holds between two facts, in which case a final position subordination should be used and an appropriate cue phrase, like *then*, is preferred in the NR clause.

Heuristic 6.1 : Preferences among features for relation-based coherence:

- a semantic relation $\prec$ CONJUNCT/DISJUNCT $\prec$ JOINT $\prec$ presuppositions of a relation not met
- JOINT $\prec$ CONJUNCT/DISJUNCT inside a semantic relation unless the relation holds for all conjoined facts

Heuristic 6.2 : Preferences among center transitions:

- Continuation $\prec$ Associate shifting $\prec$ Retaining $\prec$ Smooth shifting $\prec$ Abrupt shifting

Heuristic 6.3 : Preferences among semantic relations and center transitions:

- a semantic relation + Abrupt shifting $\prec$ JOINT + Continuation

Heuristic 6.4 : Preferences among features for embedding and center transition:

- Good embedding $\prec$ Normal embedding $\prec$ JOINT $\prec$ Bad embedding
- Continuation + Smooth shifting + JOINT $\prec$ Abrupt shifting + Normal embedding
- Good embedding $\prec$ Continuation + JOINT
- CONJUNCT $\prec$ Good embedding

A.2 A Decision Tree for Modifier Realisation

The following represents the complete decision tree for realising a given property of a discourse entity as an modifier of the NP denoting this entity. The non-terminal nodes are conditions concerning the input features: the semantic and pragmatic features of the property and the syntactic form of the NP head. The terminal nodes specify *TYPE* values and the probabilities of choosing these values. This tree is slightly different from the output of the *wagon* training program. For readability, we miss out the *TYPE* values with zero probability and sort the syntactic forms in descending probability. The form with the highest probability is chosen as the realisation of the property. When several values are equally good, the one specified first in the *wagon* program is used.

```
((cat is poss_np)
 (((poss 1) poss))
 ((sem is identify2)
  ((pragm is attr)
   (((appos .75) (such .153) (postpart .057) (postprep .019) (postnp .019)
     appos))
   ((cat is the_pn)
    (((appos .75) (preadj .222) (prenoun .027) appos))
    ((pragm is int)
     ((cat is a_np)
      (((postprep .333) (postpart .333) (rel_cls .333) postprep))
      (((such .333) (rel_cls .333) (appos .166) (postprep .166) such)))
     ((cat is bare_np)
      (((preadj .272) (postprep .272) (rel_cls .181) (appos .090)
        (prenoun .090) (postpart .090) preadj))
```

```
    ((cat is pn)
     (((preadj .666) (appos .333) preadj))
     ((cat is a_np)
      (((appos .333) (such .333) (postprep .333) appos))
      ((cat is the_np)
       (((postprep .317) (appos .268) (preadj .170) (prenoun .121)
         (such .073) (postpart .024) (rel_cls .024) postprep))
       (((appos .333) (preadj .333) (postprep .333) appos)))))))))
((sem is possinv)
 ((cat is the_np)
  (((postprep .926) (prenoun .042) (preadj .021) (poss .010) postprep))
  ((pragm is int)
   (((postprep .967) (prenoun .032) postprep))
   ((pragm is unique)
    ((cat is bare_np)
     (((postprep .625) (postpart .25) (poss .125) postprep))
     ((cat is this_np)
      (((postprep 1) postprep))
      (((poss .5) (postprep .375) (prenoun .125) poss))))
    ((cat is the_pn)
     (((preadj .666) (postprep .333) preadj))
     (((prenoun .5) (poss .25) (postprep .25) prenoun))))))
 ((sem is quality1)
  (((preadj .972) (rel_cls .020) (postprep .006) preadj))
  ((sem is characterize1)
   (((appos .806) (rel_cls .161) (prenoun .032) appos))
   ((sem is material1)
    ((cat is the_np)
     (((prenoun .842) (preadj .157) prenoun))
     ((cat is bare_np)
      (((prenoun .755) (preadj .122) (postprep .102) (postpart .020)
        prenoun))
      ((pragm is attr)
       (((prenoun .722) (postprep .166) (postpart .111) prenoun))
       ((cat is a_np)
        (((postprep .666) (prenoun .333) postprep))
        (((preadj .5) (postpart .333) (postprep .166) preadj))))))
    ((sem is property2)
     (((preadj 1) preadj))
     ((sem is location1)
      ((cat is pn)
       ((pragm is unique)
        (((postprep .5) (postnp .333) (preadj .166) postprep))
        (((postnp .571) (postpart .285) (postprep .142) postnp)))
       ((pragm is attr)
        ((cat is another_np)
         (((preadj .5) (postprep .25) (postnp .25) preadj))
         (((postprep .746) (preadj .112) (postpart .084) (prenoun .028)
           (rel_cls .028) postprep)))
        ((pragm is unique)
         ((cat is the_pn)
          (((prenoun .333) (postprep .333) (preadj .222) (postnp .111)
            prenoun))
          ((cat is bare_np)
```

```
      (((preadj .4) (postprep .4) (prenoun .12) (postpart .08) preadj))
      (((postprep .525) (preadj .254) (prenoun .135) (postpart .084)
        postprep))))
    ((cat is bare_np)
     (((postprep 1) postprep))
     ((cat is a_np)
      (((postprep .857) (prenoun .142) postprep))
      ((cat is q_np)
       (((preadj .333) (prenoun .333) (postprep .333) preadj))
       (((prenoun .428) (postprep .428) (preadj .142) prenoun))))))))
((sem is object3)
 ((cat is the_np)
  (((postprep .95) (rel_cls .05) postprep))
  ((cat is a_np)
   (((postprep 1) postprep))
   ((pragm is unique)
    (((postprep .9) (prenoun .1) postprep))
    ((cat is bare_np)
     (((postprep .75) (prenoun .25) postprep))
     (((prenoun 1) prenoun))))))
 ((sem is possess)
  ((pragm is unique)
   ((cat is the_np)
    (((postprep .571) (prenoun .285) (postpart .142) postprep))
    (((preadj .375) (postpart .375) (postprep .25) preadj)))
   ((cat is bare_np)
    ((pragm is int)
     (((postprep .333) (postpart .333) (rel_cls .333) postprep))
     (((postpart .529) (postprep .411) (prenoun .058) postpart)))
    ((pragm is int)
     ((cat is a_np)
      (((postprep .8) (postpart .2) postprep))
      (((postpart .5) (rel_cls .333) (postprep .166) postpart)))
     ((cat is the_pn)
      (((rel_cls .666) (postprep .333) rel_cls))
      ((cat is a_np)
       (((postpart .333) (rel_cls .333) (preadj .166) (prenoun .083)
         (postprep .083) postpart))
       ((cat is the_np)
        (((postprep .333) (postpart .444) (preadj .111)
          (rel_cls .111) postpart))
        (((postprep .333) (postpart .333) (rel_cls .333)
          postprep)))))))))
  ((sem is rephrase1)
   (((appos 1) appos))
   ((sem is subject7)
    ((cat is bare_np)
     (((postprep .75) (preadj .25) postprep))
     ((cat is the_np)
      (((postprep .8) (poss .066) (prenoun .066) (postpart .066)
        postprep))
      ((pragm is unique)
       (((poss .666) (postprep .333) poss))
       (((postprep 1) postprep)))))))
```

```
((sem is time_period1)
 ((pragm is unique)
  ((cat is bare_np)
   (((preadj .565) (postprep .217) (prenoun .130) (postpart .043)
     (rel_cls .043) preadj))
   ((cat is the_np)
    (((postprep .526) (preadj .315) (prenoun .052) (postpart .052)
      (postnp .052) postprep))
    ((cat is pn)
     (((postnp .75) (postprep .25) postnp))
     ((cat is the_pn)
      (((postnp .666) (preadj .333) postnp))
      (((preadj .666) (postnp .333) preadj))))))
  ((pragm is int)
   (((postprep .625) (preadj .125) (postpart .125) (rel_cls .125)
     postprep))
   ((cat is bare_np)
    (((postpart .375) (preadj .125) (prenoun .125) (postnp .125)
      (rel_cls .25) postpart))
    ((cat is the_pn)
     (((postpart .5) (rel_cls .25) (prenoun .125) (postprep .125)
       postpart))
     ((cat is a_np)
      (((postpart .428) (prenoun .285) (preadj .142) (postprep .142)
        postpart))
      ((cat is the_np)
       (((postpart .4) (preadj .2) (prenoun .2) (postprep .2)
         postpart))
       (((postprep .5) (postpart .25) (rel_cls .25) postprep)))))))))
 ((sem is content2)
  ((pragm is attr)
   (((postprep .5) (prenoun .25) (rel_cls .25) postprep))
   ((pragm is unique)
    ((cat is bare_np)
     (((preadj .6) (postprep .4) preadj))
     (((postprep .7) (postpart .2) (preadj .1) postprep)))
    (((postprep .818) (preadj .090) (other .090) postprep))))
  ((sem is purpose2)
   ((cat is num_np)
    (((postprep .666) (preadj .333) postprep))
    ((pragm is attr)
     ((cat is bare_np)
      (((preadj .5) (prenoun .25) (rel_cls .25) preadj))
      (((rel_cls .5) (prenoun .333) (postpart .166) rel_cls)))
     ((cat is bare_np)
      (((prenoun .434) (preadj .173) (postprep .173) (rel_cls .173)
        (postpart .043) prenoun))
      ((cat is the_np)
       (((preadj .357) (prenoun .285) (postprep .142) (rel_cls .142)
         (postpart .071) preadj))
       ((pragm is int)
        (((preadj .6) (prenoun .2) (postprep .1) (postpart .1)
          preadj))
        ((cat is poss_pro)
```

```
      (((preadj .25) (prenoun .25) (postprep .25) (postpart .25)
        preadj))
      (((postprep .5) (preadj .25) (postpart .25) postprep))))))))
((sem is other)
 ((pragm is attr)
  ((cat is a_np)
   (((preadj .4) (postpart .3) (prenoun .1) (rel_cls .1)
     (other .1) preadj))
   ((cat is the_pn)
    (((rel_cls .666) (postpart .333) rel_cls))
    ((cat is bare_np)
     (((preadj .357) (postpart .285) (rel_cls .285)
       (postprep .071) preadj))
     ((cat is the_np)
      (((preadj .5) (postpart .25) (rel_cls .25) preadj))
      (((rel_cls .5) (prenoun .25) (postpart .25) rel_cls))))))
  (((preadj .633) (rel_cls .116) (postprep .108) (postpart .091)
    (prenoun .05) preadj)))
 ((sem is unsureSEM)
  ((pragm is attr)
   (((other .666) (rel_cls .333) other))
   ((cat is bare_np)
    ((pragm is unique)
     (((such .333) (preadj .333) (prenoun .333) such))
     (((preadj .666) (postpart .333) preadj)))
    (((postprep .666) (preadj .333) postprep))))
  ((sem is temporal_property1)
   (((preadj .974) (postnp .025) preadj))
   ((sem is state4)
    (((preadj 1) preadj))
    ((cat is the_np)
     ((sem is visual_property1)
      ((pragm is attr)
       (((rel_cls .5) (preadj .25) (prenoun .25) rel_cls))
       (((preadj .833) (prenoun .166) preadj)))
      ((sem is spatial_property1)
       ((pragm is unique)
        (((prenoun .5) (preadj .25) (postprep .25) prenoun))
        (((preadj 1) preadj)))
       (((preadj 1) preadj))))
     (((preadj .762) (prenoun .125) (postpart .075)
       (postprep .025) (postnp .012) preadj))))))))))))))))))))))))
```

A.3 Adjective Ordering

In our domain, a head noun often needs to be modified by multiple adjectives, such as *the important Scottish designer Jessie King*, so premodifier ordering is an indispensable task. Based on previous studies of adjective ordering (Abberton, 1977; Dale, 1992), we use the following ordering scheme in our implementations to order prehead modifiers, and the scheme is generalised using the concepts of the Generalized Upper-Model (only adjectives for material world qualities are considered):

`g s a c p o & N H`

The meaning of the symbols are:

g : Evaluative-quality of moral, aesthetic, or utilitarian, e.g. *honest, beautiful, readable.*

s, a, c : Sense-and-measure-quality of size, age and colour, e.g. *big, young, red, expensive.*

p : Status-quality and Behavioral-quality, including participles, e.g. *empty, skillful, dying.*

o : origin, i.e. Class-quality including Provenance-class-quality and Material-class-quality, e.g. *British, wooden.*

& : coordinating items as *and, or, but, yet.*

N : Prehead nominals including genitival nouns, proper names, and group genitives.

H : head noun.

There is also work on corpus based adjective ordering, e.g. (Shaw and Hatzivassiloglou, 1999), which could be a direction for future work.

Appendix B

Questionnaires

This appendix presents the questionnaires we used for our experiments with human subjects. The page breaks are changed to save space.

B.1 Assessing Similarities between Constructions

This questionnaire is designed for you to assess the similarities between sentences and the naturalness of sentences. In each group, there is one leading sentence signalled by boldface font in its context. One or two following sentences are given to be compared with the leading sentence in the same context. Please read them carefully and circle the appropriate number that indicates your assessment of:

- the degree of *similarity* in meaning between a subsequent sentence and its leading sentence according to the following rating:

  6 - exactly the same
  5 - very similar
  4 - more similar than different
  3 - more different than similar
  2 - very different
  1 - totally different

- the degree of *naturalness* of the same sentence using the following rating:

  5 - natural
  4 - fairly natural
  3 - so-so
  2 - fairly unnatural
  1 - unnatural

Here is an example. You will be given a group of sentences as follows:

1. The buy-out group's task of holding its fragile coalition together has been further complicated by an apparent rift in the ranks of the pilot union itself. A pilot representing a

group of 220 pilots filed suit Friday in Chicago federal court to block the takeover. **The dissident pilots oppose the plan because it would cause them to lose their seniority.**

(a) The dissident pilots oppose the plan, which would cause them to lose their seniority.

Similarity :	6	5	[4]	3	2	1
Naturalness:	[5]	4	3	2	1	

(b) The dissident pilots oppose the plan. As for the plan, it would cause them to lose their seniority.

Similarity :	6	5	4	3	2	[1]
Naturalness:	5	4	[3]	2	1	

First you compare sentence (a) with sentence 1 and circle the number you think that indicates their similarity (here 4 is chosen for example), and then you rate the naturalness of sentence (a) (in the example it is 5). For (b), you do exactly the same.

Now you have completed the instructions. The questionnaire starts next.

1. SFE Technologies said William P. Kuehn was elected chairman and chief executive officer of this troubled electronics parts maker. **The 45-year-old Mr. Kuehn, who has a background in crisis management, succeeds Alan D. Rubendall, 45.**

   (a) The 45-year-old Mr. Kuehn succeeds Alan D. Rubendall, 45. As for Mr. Kuehn, he has a background in crisis management.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) The 45-year-old Mr. Kuehn succeeds Alan D. Rubendall, 45 because he has a background in crisis management.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

2. As financial markets rebounded, trading volume in the Chicago Mercantile Exchange's huge Standard & Poor's 500 stock-index futures pit soared, reaching near-record levels for the first time since October 1987. **The S&P 500 futures contract, which jumped two to three points in seconds early yesterday after an initial downturn, moved strongly higher the rest of the day.**

   (a) The S&P 500 futures contract jumped two to three points in seconds early yesterday after an initial downturn, and then moved strongly higher the rest of the day.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) The S&P 500 futures contract moved strongly higher the rest of yesterday. What is more, it jumped two to three points in seconds early yesterday after an initial downturn.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

3. After the collapse of the last effort, the group doesn't plan to make any formal proposal without binding commitments from banks covering the entire amount to be borrowed. **Under the type of transaction being discussed, the pilot-management group would borrow from banks several billion dollars, which could be used to finance a cash payment to current holders.**

   (a) Under the type of transaction being discussed, the pilot-management group would borrow from banks several billion dollars. As for the money, it could be used to finance a cash payment to current holders

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) Under the type of transaction being discussed, the pilot-management group would borrow from banks several billion dollars and the money could then be used to finance a cash payment to current holders.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

4. Eastern and its creditors agreed in July on a reorganization plan that called for the carrier to sell off $1.8 billion in assets and to emerge from Chapter 11 status in late 1989 at two-thirds its former size. **Eastern eventually decided not to sell off a major chunk, its South American routes, which were valued at $400 million.**

   (a) Eastern eventually decided not to sell off a major chunk, its South American routes, because they were valued at $400 million.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) Eastern eventually decided not to sell off a major chunk, its South American routes. As for the routes, they were valued at $400 million.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

5. New York financier Saul Steinberg sought federal permission to buy more than 15% of United Airlines' parent, UAL Corp., saying he might seek control of the nation's second-largest airline. **But any potential acquirer must attempt to reach some kind of accord with the company's employees, primarily its pilots and the powerful machinists' union, which has opposed a takeover.**

   (a) But any potential acquirer must attempt to reach some kind of accord with the company's employees, primarily its pilots and the powerful machinists' union because it has opposed a takeover.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) But any potential acquirer must attempt to reach some kind of accord with the company's employees, primarily its pilots and the powerful machinists' union. As for the union, it has opposed a takeover.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

6. A REVISED BID FOR UAL is being prepared by a labor-management group, sources said. **The new proposal, which would transfer majority ownership of United Air's parent to employees and leave some stock in public hands, would be valued at as much as $5.42 billion.**

   (a) The new proposal would be valued at as much as $5.42 billion because it would transfer majority ownership of United Air's parent to employees and leave some stock in public hands.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

   (b) The new proposal would be valued at as much as $5.42 billion. As for the proposal, it would transfer majority ownership of United Air's parent to employees and leave some stock in public hands.

Similarity : 6 5 4 3 2 1
Naturalness: 5 4 3 2 1

7. Mr. Bianchi said some big European investors were selling dollars in lots of $100 million to $200 million, which led to nervousness in the trading room. Yet Heiko Thieme, an investment strategist for Deutsche Bank in New York, contended that Europeans hadn't purchased many American shares this year and the dollar wasn't vulnerable at all. **Mr Thieme said that on a fundamental basis, he was not afraid about the dollar, which ran more of a risk of being too strong than too weak.**

   (a) Mr Thieme said that on a fundamental basis, he was not afraid about the dollar because it ran more of a risk of being too strong than too weak.

   Similarity : 6 5 4 3 2 1
   Naturalness: 5 4 3 2 1

   (b) Mr Thieme said that on a fundamental basis, he was not afraid about the dollar, as for the dollar, it ran more of a risk of being too strong than too weak.

   Similarity : 6 5 4 3 2 1
   Naturalness: 5 4 3 2 1

8. Although final details weren't available, sources said the Dingell plan would abandon the president's proposal for a cap on utilities' sulfur-dioxide emissions. **That proposal had been hailed by environmentalists but despised by utilities because they feared it would limit their growth.**

   (a) That proposal had been hailed by environmentalists but despised by utilities, who feared it would limit their growth.

   Similarity : 6 5 4 3 2 1
   Naturalness: 5 4 3 2 1

9. Spiegel said margins improved because its inventory position this year didn't need the costly markdowns required to trim last year's swollen levels. **A spokeswoman said the apparel market troughed in the first half of 1988, then began showing improvement in the second half of that year.**

   (a) A spokeswoman said the apparel market, which troughed in the first half of 1988, then began showing improvement in the second half of that year.

   Similarity : 6 5 4 3 2 1
   Naturalness: 5 4 3 2 1

   (b) A spokeswoman said the apparel market, which troughed in the first half of 1988, began showing improvement in the second half of that year.

   Similarity : 6 5 4 3 2 1
   Naturalness: 5 4 3 2 1

10. Mr. McGovern, 63, had been under intense pressure from the board to boost Campbell's mediocre performance to the level of other food companies. **The board is dominated by the heirs of the late John T. Dorrance Jr., who controlled about 58% of Campbell's stock when he died in April.**

(a) The board is dominated by the heirs of the late John T. Dorrance Jr. because he controlled about 58% of Campbell's stock when he died in April.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) The board is dominated by the heirs of the late John T. Dorrance Jr. As for Mr. Dorrance, he controlled about 58% of Campbell's stock when he died in April.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

11. Harry Manion, Mr. Sala's attorney, says his client denies any wrongdoing and adds that the attorney general's contentions about First Meridian's business practices are incorrect. As for Mr. Sala's qualifications, Mr. Manion says the snooty attorneys for the state of New York decided **Mr. Sala wasn't qualified because he didn't have a Harvard degree.**

(a) Mr. Sala, who didn't have a Harvard degree, wasn't qualified.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

12. National Medical Enterprises Inc. said the completion of the spinoff of its long-term care operations will be delayed until early next year because of regulatory complexities. **The health-care services concern announced the spinoff plan last January. The plan was then revised in May and was hoped to be completed by Nov. 30.**

(a) The health-care services concern announced the spinoff plan last January, which was revised in May and hoped to be completed by Nov. 30.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) The health-care services concern announced the spinoff plan last January, which was then revised in May and hoped to be completed by Nov. 30.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

13. Over at the fiberglass factory, four white workers assemble water tanks on their own, and in their spare time they build townhouses across the road. **On Main Street, Alida Verwoerd and her daughters look after the clothes and fabric shop, then hurry home to fix lunch for the rest of the family.**

(a) On Main Street, Alida Verwoerd and her daughters, who first look after the clothes and fabric shop, then hurry home to fix lunch for the rest of the family.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) On Main Street, Alida Verwoerd and her daughters, who look after the clothes and fabric shop, then hurry home to fix lunch for the rest of the family.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

14. Founded as the Examiner in 1903 by Mr. Hearst, the Herald was crippled by a bitter, decade-long strike that began in 1967 and cut circulation in half. Financially, it never recovered; editorially, it had its moments. **In 1979, Hearst hired editor James Bellows, who brightened the editorial product considerably.**

(a) In 1979, Hearst hired editor James Bellows, and he then brightened the editorial product considerably.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) In 1979, Hearst hired editor James Bellows. As for Mr. Bellows, he brightened the editorial product considerably.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

15. Eastern Reproduction Corp., maker of thin metal precision parts, must report to five federal and state agencies as well as to local fire, police, hospital and plumbing authorities. **One state environmental regulator returned a report because it wasn't heavy enough, Mr. Maguire says.**

(a) One state environmental regulator returned a report, which wasn't heavy enough, Mr. Maguire says.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

16. The adjustments result from the recently passed thrift-industry bailout legislation, which requires thrifts to divest all high-yield bond investments by 1994. **Columbia didn't have to adjust the book value of its junk-bond holdings to reflect declines in market prices, because it held the bonds as long-term investments.**

(a) Previously, Columbia, which held the bonds as long-term investments, didn't have to adjust the book value of its junk-bond holdings to reflect declines in market prices.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

17. In 1953, James Watson and his colleagues unlocked the double helix of DNA, the genetic key to heredity. **Twenty years later, two California academics made "recombinant" DNA, transplanting a toad's gene into bacteria, which reproduced toad genes.**

(a) Twenty years later, two California academics made "recombinant" DNA, transplanting a toad's gene into bacteria. What is more, the bacteria reproduced toad genes.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) Twenty years later, two California academics made "recombinant" DNA, transplanting a toad's gene into bacteria, and the bacteria then reproduced toad genes.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

18. **Italy should close the Leaning Tower of Pisa because it's a danger to tourists, government-appointed experts said.** "In some places the stonework is so damaged it shows signs of breaking off," scientists and technicians said in a report to Public Works Minister Giovanni Prandini.

(a) Italy should close the Leaning Tower of Pisa, which is a danger to tourists, government-appointed experts said.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

19. Insurance companies have been major buyers of prepayment-protected planned amortization classes (PACs) during the past few months. **The PACs, which have higher yields than topgrade corporate bonds, appeal to insurance companies and other investors.**

(a) The PACs appeal to insurance companies and other investors because they have higher yields than topgrade corporate bonds.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) The PACs appeal to insurance companies and other investors. What is more, they have higher yields than topgrade corporate bonds.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

20. The British satirical magazine Private Eye won an appeal against the size of a $960,000 libel award to Sonia Sutcliffe, the estranged wife of the "Yorkshire Ripper" mass murderer. An appeals-court panel slashed all but $40,000 from the award, the largest ever set by a British jury. **Private Eye had been threatened with closure because it couldn't afford the libel payment.**

(a) Private Eye, which couldn't afford the libel payment, had been threatened with closure.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) Private Eye, which had been threatened with closure, couldn't afford the libel payment.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

21. But P&G contends the new Cheer is a unique formula that also offers an ingredient that prevents colors from fading. **And retailers are expected to embrace the product, because it will take up less shelf space.**

(a) And retailers are expected to embrace the product, which will take up less shelf space.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

22. The earthquake rocked northern California last week. When Aetna adjuster Bill Schaeffer visited a retired couple in Oakland last Thursday, he found them living in a mobile home parked in front of their yard. **Their house, which was pushed about four feet off its foundation, collapsed into its basement.**

(a) Their house was collapsed into its basement. What is more, it was pushed about four feet off its foundation.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) Their house was pushed about four feet off its foundation, and then collapsed into its basement.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

23. J.L. Henry & Co., Miami, and a principal of the firm, Henry Otero of Miami, were jointly fined $30,000 and expelled, for alleged improper use of a customer's funds, among other things. J.L. Henry hasn't any Miami telephone listing, an operator said. **Mr. Otero, who apparently has an unpublished number, also couldn't be reached.**

(a) Mr. Otero also couldn't be reached. What is more, he apparently has an unpublished number.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) Mr. Otero also couldn't be reached because he apparently has an unpublished number.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

24. Medical researchers believe the transplantation of small amounts of fetal tissue into humans could help treat juvenile diabetes and some degenerative diseases. **But anti-abortionists oppose such research because they worry that the development of therapies using fetal-tissue transplants could lead to an increase in abortions.**

(a) But anti-abortionists, who worry that the development of therapies using fetal-tissue transplants could lead to an increase in abortions, oppose such research.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

(b) But anti-abortionists, who oppose such research, worry that the development of therapies using fetal-tissue transplants could lead to an increase in abortions.

Similarity :	6	5	4	3	2	1
Naturalness:	5	4	3	2	1	

Thank you very much for filling in this questionnaire!

B.2 Judging Text Coherence

In this experiment, you will be given a number of short texts describing jewels that you may see in a museum. You should judge the coherence or fluency of each text by assigning a score to it.

Coherence or fluency in this experiment is defined as how well you think the text is arranged, and perhaps how it flows. You might find that some texts are more interesting than others, or feel that repetitions between texts irritating. But please don't let these or differences in the use of vocabulary bias your judgment of the fluency of each individual text.

The coherence of each text is to be scored using a number between 0 and 10, where a higher number represents a more fluent text. On each page are three descriptions of the same jewel and there are 6 different jewels in this experiment. Please read each text carefully and then score the coherence of the text by circle the number that indicates your assessment. For example, you might want to use the number 5 for a text that you think is OK or so-so, and then use higher or lower numbers for those texts that you feel positive or negative toward their coherence. You might also want to judge a text by comparing it with the other two texts on the same page.

We are interested in your first impressions, so do concentrate but please don't take too much time to think about any one text: try to make up your mind quickly, spending less than a minute on each text. If you have further comments about the texts, please put them on the last sheet.

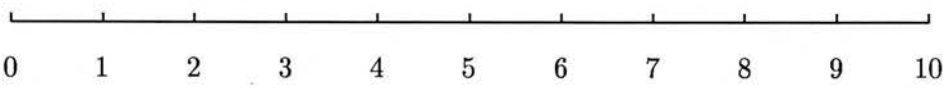
THANK YOU VERY MUCH FOR YOUR TIME!

1. This jewel is set with jewels. It is a necklace and was made in 1910. The jewel is made from turquoise, mother-of-pearl, silver metal, glass, beryl and tourmalines. It uses oval-shaped stones. In other words, it features rounded stones. The jewel was made by Arthur and Georgie Gaskin. It is in the Arts-and-Crafts style and has an elaborate design. It was produced by single craftsman.

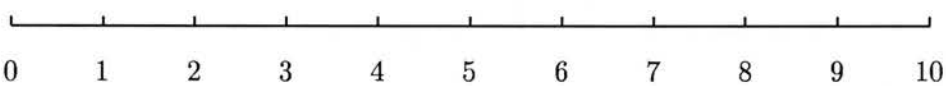
The Gaskins are British. They are important and lived in Birmingham.

Arts and Crafts style jewels usually demonstrate the artistic sensibilities of the wearer. They usually use oval-shaped stones and usually feature rounded stones. They were usually produced by single craftsman. This jewel uses natural objects with imperfections in that it incorporates flawed stones.

Arts and Crafts style jewels usually have an elaborate design and are usually flexible.



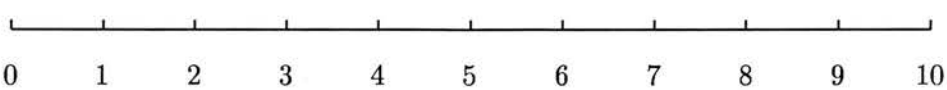
2. This jewel is a necklace and was made by the important British designers Arthur and Georgie Gaskin, who lived in Birmingham. It is in the Arts and Crafts style and was made in 1910. The jewel is made from silver metal, beryl, tourmalines, turquoise, mother-of-pearl and glass. It is set with jewels. The jewel features rounded stones. Indeed Arts and Crafts style jewels usually feature rounded stones; for instance this jewel uses oval-shaped stones (indeed Arts and Crafts style jewels usually use oval-shaped stones). Like most Arts and Crafts style jewels, this jewel was produced by single craftsman. It uses natural objects with imperfections, in that it incorporates flawed stones. It has an elaborate design; indeed Arts and Crafts style jewels usually have an elaborate design. They are usually flexible and usually demonstrate the artistic sensibilities of the wearer.



3. This jewel uses natural objects with imperfections in that it incorporates flawed stones. It was produced by single craftsman.

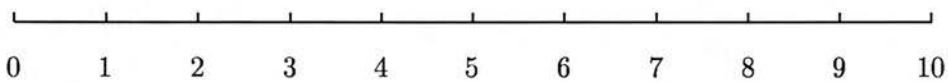
The jewel, which was made in 1910, was made by the important British designers Arthur and Georgie Gaskin, who lived in Birmingham. It is in the Arts-and-Crafts style and is a necklace. The jewel is set with jewels in that it features rounded stones. Indeed, Arts and Crafts style jewels usually feature rounded stones. For example this jewel uses oval-shaped stones. Indeed, Arts and Crafts style jewels usually use oval-shaped stones.

This jewel is made from beryl, turquoise, tourmalines, glass, silver metal and mother-of-pearl. It has an elaborate design. Indeed, Arts and Crafts style jewels usually have an elaborate design. They usually demonstrate the artistic sensibilities of the wearer and are usually flexible. They are usually produced by single craftsman.



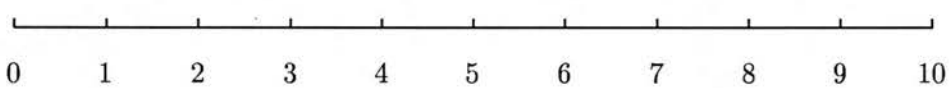
4. This jewel, which is 46.50 cm in length, is a necklace. It draws for inspiration on machines and their components. It is made from plastic and polished steel.

The jewel, which was made in England, was made in 1920. It is in the Machine-age style. The jewel has no fear of pattern in that it incorporates patterns with a repetitive element. Indeed, Machine-age style jewels usually incorporate patterns with a repetitive element. For example this jewel has regularly repeated forms in that it is made up of a pattern of interlocking rods. Indeed, Machine-age style jewels usually have regularly repeated forms.

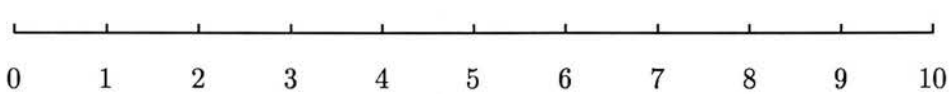


5. This jewel is in the Machine-age style. It draws for inspiration on machines and their components. The jewel is 46.50 cm in length and is a necklace. It has no fear of pattern in that it incorporates patterns with a repetitive element. Indeed, Machine-age style jewels usually incorporate patterns with a repetitive element. For example this jewel is made up of a pattern of interlocking rods. Indeed, Machine-age style jewels usually have regularly repeated forms.

This jewel is made from polished steel and plastic. It was made in 1920 and has regularly repeated forms. It was made in England.

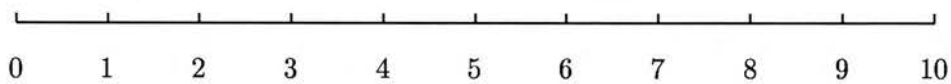


6. This jewel is a necklace, which is 46.50 cm in length. It is in the machine-age style. The jewel was made in 1920. It is made from plastic and polished steel. The jewel was made in England and draws for inspiration on machines and their components. It has no fear of pattern, in that it incorporates patterns with a repetitive element. Machine-age style jewels usually have regularly repeated forms. To take an example: this jewel is made up of a pattern of interlocking rods; indeed machine-age style jewels usually incorporate patterns with a repetitive element (for instance this jewel has regularly repeated forms).



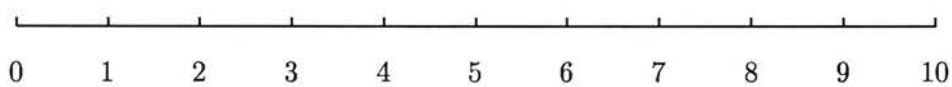
7. This jewel is a finger ring and was made by the British designer Ernest Blyth, who lived in the United Kingdom. It is in the Organic style and was made in 1968. The jewel is made from 18-carat gold, aquamarines and diamonds. It was made in the United Kingdom and has a coarse texture. The jewel draws on natural themes for inspiration. It is set with jewels. The jewel is encrusted with gems, in that it has little diamonds scattered around its edges; indeed Organic style jewels are usually encrusted with gems.

Organic style jewels are usually made up of asymmetrical shapes. They usually have a coarse texture. To take an example: this jewel has heavily-textured gold; indeed Organic style jewels usually draw on natural themes for inspiration (for instance this jewel looks crystalline).



8. This jewel is made from diamonds, aquamarines and 18-carat gold. It is a finger ring and is in the Organic style. The jewel is set with jewels in that it is encrusted with gems. Indeed, Organic style jewels are usually encrusted with gems. For example this jewel has little diamonds scattered around its edges. It draws on natural themes for inspiration in that it looks crystalline. Indeed, Organic style jewels usually draw on natural themes for inspiration.

This jewel, which was made in the United Kingdom, was made in 1968. It was made by the British designer Ernest Blyth, who lived in the United Kingdom. The jewel has a coarse texture in that it has heavily-textured gold. Indeed, Organic style jewels usually have a coarse texture. They are usually made up of asymmetrical shapes.

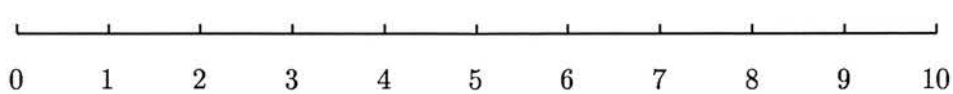


9. This jewel is set with jewels. It is in the Organic style. The jewel looks crystalline. Indeed, Organic style jewels usually draw on natural themes for inspiration. For example this jewel has heavily-textured gold. Indeed, Organic style jewels usually have a coarse texture.

This jewel, which has a coarse texture, was made in the United Kingdom. It was made by the British designer Ernest Blyth, who lived in the United Kingdom. The jewel is made from diamonds, aquamarines and 18-carat gold. It is a finger ring.

Organic style jewels are usually made up of asymmetrical shapes.

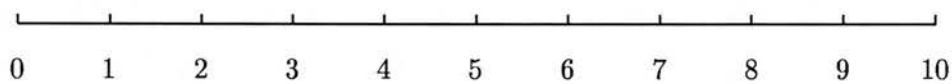
This jewel was made in 1968 and draws on natural themes for inspiration. It is encrusted with gems in that it has little diamonds scattered around its edges. Indeed, Organic style jewels are usually encrusted with gems.



10. This jewel is a necklace. It is made from gold, enamel and sapphire. The jewel was made in London. It is set with jewels in that it uses faceted stones, although it is in the Arts-and-Crafts style.

The jewel, which was made in 1905, was made by the important Scottish designer Jessie M. King, who lived in London. It was made for the British company Liberty and Co, which is based in Regent St., London. The jewel was produced in limited quantity and has festoons. It has an elaborate design; specifically, it has floral motifs. Indeed, Arts and Crafts style jewels usually have an elaborate design.

Arts and Crafts style jewels usually feature rounded stones. They usually demonstrate the artistic sensibilities of the wearer and are usually flexible. They usually use oval-shaped stones and were usually produced by single craftsman.

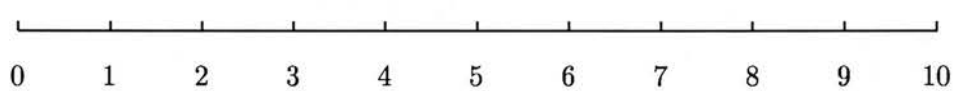


11. This jewel is in the Arts-and-Crafts style. It was made in London. The jewel is set with jewels in that it uses faceted stones, although Arts and Crafts style jewels usually use oval-shaped stones. This jewel was made for the British company Liberty and Co, which is based in Regent St., London. It is made from sapphire, enamel and gold.

The important designer Jessie M. King is Scottish. She lived in London.

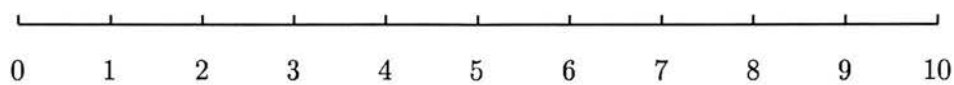
This jewel is a necklace. It was made by King and was produced in limited quantity. The jewel has an elaborate design; specifically, it has floral motifs. It has festoons and was made in 1905.

Arts and Crafts style jewels usually have an elaborate design and usually demonstrate the artistic sensibilities of the wearer. Arts and Crafts style jewels, which were usually produced by single craftsman, are usually flexible. They usually feature rounded stones.

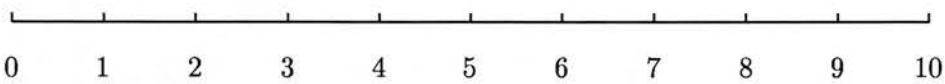


12. This jewel is a necklace and is in the Arts and Crafts style. It was made by the important British designer Jessie M. King, who lived in London. The jewel was made in 1905. It is made from gold, enamel and sapphire. The jewel was made in London and has festoons. It was produced in limited quantity and was made for Liberty and Co, which is based in Regent St., London. It is set with jewels. Although Arts and Crafts style jewels usually feature rounded stones it uses faceted stones.

Arts and Crafts style jewels usually use oval-shaped stones and are usually flexible. They usually demonstrate the artistic sensibilities of the wearer and were usually produced by single craftsman. They usually have an elaborate design (for instance this jewel has floral motifs).



13. This jewel is a necklace, which is 72.00 cm in length. It was made by the British designer Gerda Flockinger, who lived in London. The jewel, which was made in London, is in the Organic style and was made in 1976. It is made from silver metal, gold, pearls, diamonds and opals. The jewel draws on natural themes for inspiration. It is 72.00 cm in length. It is set with jewels. It is encrusted with gems, in that it has silver links encrusted asymmetrically with pearls and diamonds; indeed Organic style jewels are usually encrusted with gems. They usually have a coarse texture and are usually made up of asymmetrical shapes. They usually draw on natural themes for inspiration (for instance this jewel uses natural pearls).



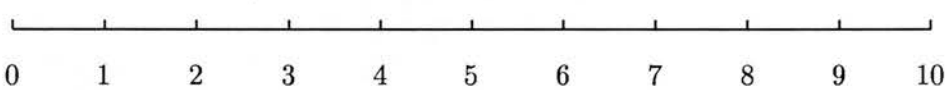
14. This jewel was made by the British designer Gerda Flockinger.

Organic style jewels are usually made up of asymmetrical shapes.

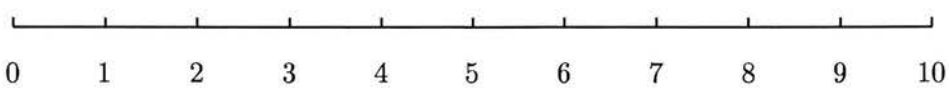
Flockinger lived in London.

This jewel has silver links encrusted asymmetrically with pearls and diamonds and is in the Organic style. It draws on natural themes for inspiration in that it uses natural pearls. Indeed, Organic style jewels usually draw on natural themes for inspiration. This jewel was made in 1976. It is set with jewels in that it is encrusted with gems. Indeed, Organic style jewels are usually encrusted with gems. They usually have a coarse texture.

This jewel was made in London and is a necklace. It is made from opals, gold, pearls, silver metal and diamonds. It is 72.00 cm in length.



15. This jewel, which is 72.00 cm in length, is a necklace. It is in the Organic style. The jewel was made by the British designer Gerda Flockinger, who lived in London. It was made in London and was made in 1976. It is set with jewels in that it is encrusted with gems. Indeed, Organic style jewels are usually encrusted with gems. For example this jewel has silver links encrusted asymmetrically with pearls and diamonds. It is made from silver metal, diamonds, gold, pearls and opals. The jewel draws on natural themes for inspiration in that it uses natural pearls. Indeed, Organic style jewels usually draw on natural themes for inspiration. They are usually made up of asymmetrical shapes and usually have a coarse texture.



16. Organic style jewels usually draw on natural themes for inspiration and usually have a coarse texture. They are usually made up of asymmetrical shapes and is usually encrusted with gems.

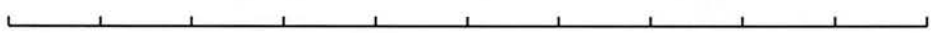
This jewel is 37 mm long and 32 mm wide. It is encrusted with gems in that it features diamonds encrusted on a natural shell. The jewel was made in 1973. It is in the Organic style and is inscribed with Hallmarks: JKM in trefoil. Crown. 18. London Assay Office mark. Date. Letter s(1973). It was made by the British designer Jacqueline Mina, who lived in London. The jewel is set with jewels. It was made in London and is a finger ring. It is made from 18-carat gold, shell and diamonds.

0 1 2 3 4 5 6 7 8 9 10

17. This jewel is 37 mm long and 32 mm wide. It is made from diamonds, 18-carat gold and shell. It is in the Organic style.

The jewel, which was made in 1973, was made in London. It is a finger ring and is inscribed with Hallmarks: JKM in trefoil. Crown. 18. London Assay Office mark. Date. Letter s(1973). The jewel was made by the British designer Jacqueline Mina, who lived in London. It is set with jewels in that it is encrusted with gems. Indeed, Organic style jewels are usually encrusted with gems. For example this jewel features diamonds encrusted on a natural shell.

Organic style jewels usually have a coarse texture. They are usually made up of asymmetrical shapes and usually draw on natural themes for inspiration.



0 1 2 3 4 5 6 7 8 9 10

18. This jewel is a finger ring, which is 37 mm long. It was made by the British designer Jacqueline Mina, who lived in London. The jewel, which is inscribed with Hallmarks: JKM in trefoil. Crown. 1 8. London Assay Office mark. Date. Letter s(1973), is in the Organic style. It was made in 1973. The jewel is made from 18-carat gold, shell and diamonds. It was made in London. The jewel is 32 mm wide and 37 mm long. It is set with jewels. It is encrusted with gems, in that it features diamonds encrusted on a natural shell; indeed Organic style jewels are usually encrusted with gems. Organic style jewels usually have a coarse texture and are usually made up of asymmetrical shapes. They usually draw on natural themes for inspiration.

0 1 2 3 4 5 6 7 8 9 10

Comments (please put any remark about the texts here)

B.3 Assessing Inferrability

This questionnaire is designed for you to assess the inferrability between two sentences. By *Inferrability* we mean:

> Given two separate sentences/clauses, the likelihood of you inferring from your world knowledge that a causal/temporal connection between the sentences/clauses might plausibly exist.

In other words, according to your world knowledge how likely it is that the situation described in one sentence causes or happens before the situation described in the other.

In each test item, there are two sentences in boldface which are to be read in their context in smaller font. Please read the sentences carefully and circle the appropriate number that indicates your assessment of the inferrability between the two sentences according to the following rating:

5 - very likely
4 - quite likely
3 - possibily
2 - even less possibily
1 - don't know

The ordering between the two sentences are not important. The "**" in the test item means that the boldfaced sentences go into that position rather than the end of the context as is the case of most items. The questionnaire starts from the next page.

1. SFE Technologies said William P. Kuehn was elected chairman and chief executive officer of this troubled electronics parts maker.

   - **The 45-year-old Mr. Kuehn succeeds Alan D. Rubendall, 45.**
   - **Mr. Kuehn has a background in crisis management.**

   Inferrability : 1 2 3 4 5

2. As financial markets rebounded, trading volume in the Chicago Mercantile Exchange's huge Standard & Poor's 500 stock-index futures pit soared, reaching near-record levels for the first time since October 1987.

   - **The S&P 500 futures contract jumped two to three points in seconds early yesterday after an initial downturn.**
   - **It moved strongly higher the rest of the day.**

   Inferrability : 1 2 3 4 5

3. Mr. McGovern, 63, had been under intense pressure from the board to boost Campbell's mediocre performance to the level of other food companies.

   - **The board is dominated by the heirs of the late John T. Dorrance Jr.**
   - **He controlled about 58% of Campbell's stock when he died in April.**

   Inferrability : 1 2 3 4 5

4. Eastern and its creditors agreed in July on a reorganization plan that called for the carrier to sell off $1.8 billion in assets and to emerge from Chapter 11 status in late 1989 at two-thirds its former size.

   - **Eastern eventually decided not to sell off a major chunk, its South American routes.**
   - **The South American routes were valued at $400 million.**

   Inferrability : 1 2 3 4 5

5. After the collapse of the last effort, the group doesn't plan to make any formal proposal without binding commitments from banks covering the entire amount to be borrowed.

   - **Under the type of transaction being discussed, the pilot-management group would borrow from banks several billion dollars.**

- **This money could be used to finance a cash payment to current holders.**

Inferrability : 1 2 3 4 5

6. A REVISED BID FOR UAL is being prepared by a labor-management group, sources said.

   - **The new proposal would be valued at as much as $5.42 billion.**
   - **It would transfer majority ownership of United Air's parent to employees and leave some stock in public hands.**

   Inferrability : 1 2 3 4 5

7. Mr. Bianchi said some big European investors were selling dollars in lots of $100 million to $200 million, which led to nervousness in the trading room. Yet Heiko Thieme, an investment strategist for Deutsche Bank in New York, contended that Europeans hadn't purchased many American shares this year and the dollar wasn't vulnerable at all.

   - **Mr Thieme said that on a fundamental basis, he was not afraid about the dollar,**
   - **the dollar ran more of a risk of being too strong than too weak.**

   Inferrability : 1 2 3 4 5

8. Although final details weren't available, sources said the Dingell plan would abandon the president's proposal for a cap on utilities' sulfur-dioxide emissions.

   - **That proposal had been hailed by environmentalists but despised by utilities.**
   - **Utilities feared it would limit their growth.**

   Inferrability : 1 2 3 4 5

9. Spiegel said margins improved because its inventory position this year didn't need the costly markdowns required to trim last year's swollen levels.

   - **A spokeswoman said the apparel market troughed in the first half of 1988,**
   - **the market began showing improvement in the second half of that year.**

   Inferrability : 1 2 3 4 5

10. New York financier Saul Steinberg sought federal permission to buy more than 15% of United Airlines' parent, UAL Corp., saying he might seek control of the nation's second-largest airline.

    - **But any potential acquirer must attempt to reach some kind of accord with the company's employees, primarily its pilots and the powerful machinists' union.**
    - **The union has opposed a takeover.**

    Inferrability : 1 2 3 4 5

11. Harry Manion, Mr. Sala's attorney, says his client denies any wrongdoing and adds that the attorney general's contentions about First Meridian's business practices are incorrect. As for Mr. Sala's qualifications, Mr. Manion says the snooty attorneys for the state of New York decided

    - **Mr. Sala wasn't qualified.**
    - **He didn't have a Harvard degree.**

    Inferrability : 1 2 3 4 5

12. National Medical Enterprises Inc. said the completion of the spinoff of its long-term care operations will be delayed until early next year because of regulatory complexities.

    - **The health-care services concern announced the spinoff plan last January.**
    - **The plan was revised in May and hoped to be completed by Nov. 30.**

    Inferrability : 1 2 3 4 5

13. Over at the fiberglass factory, four white workers assemble water tanks on their own, and in their spare time they build townhouses across the road.

    - **On Main Street, Alida Verwoerd and her daughters look after the clothes and fabric shop.**
    - **They hurry home to fix lunch for the rest of the family.**

    Inferrability : 1 2 3 4 5

14. Founded as the Examiner in 1903 by Mr. Hearst, the Herald was crippled by a bitter, decade-long strike that began in 1967 and cut circulation in half. Financially, it never recovered; editorially, it had its moments.

    - **In 1979, Hearst hired editor James Bellows.**
    - **Bellows brightened the editorial product considerably.**

    Inferrability : 1 2 3 4 5

15. Eastern Reproduction Corp., maker of thin metal precision parts, must report to five federal and state agencies as well as to local fire, police, hospital and plumbing authorities.

    - **One state environmental regulator returned a report.**
    - **The report wasn't heavy enough.**

    Inferrability : 1 2 3 4 5

16. The adjustments result from the recently passed thrift-industry bailout legislation, which requires thrifts to divest all high-yield bond investments by 1994.

    - **Previously, Columbia didn't have to adjust the book value of its junk-bond holdings to reflect declines in market prices.**
    - **Columbia held the bonds as long-term investments.**

    Inferrability : 1 2 3 4 5

17. In 1953, James Watson and his colleagues unlocked the double helix of DNA, the genetic key to heredity.

    - **Twenty years later, two California academics made "recombinant" DNA, transplanting a toad's gene into bacteria.**
    - **The bacteria reproduced toad genes.**

    Inferrability : 1 2 3 4 5

18. ** "In some places the stonework is so damaged it shows signs of breaking off," scientists and technicians said in a report to Public Works Minister Giovanni Prandini.

    - **Italy should close the Leaning Tower of Pisa,**
    - **it is a danger to tourists, government-appointed experts said.**

    Inferrability : 1 2 3 4 5

19. Insurance companies have been major buyers of prepayment-protected planned amortization classes (PACs) during the past few months.

    - **The PACs appeal to insurance companies and other investors.**
    - **They have higher yields than topgrade corporate bonds.**

    Inferrability : 1 2 3 4 5

20. The British satirical magazine Private Eye won an appeal against the size of a $960,000 libel award to Sonia Sutcliffe, the estranged wife of the "Yorkshire Ripper" mass murderer. An appeals-court panel slashed all but $40,000 from the award, the largest ever set by a British jury.

    - **Private Eye had been threatened with closure.**
    - **It couldn't afford the libel payment.**

    Inferrability : 1 2 3 4 5

21. But P&G contends the new Cheer is a unique formula that also offers an ingredient that prevents colors from fading.

    - **And retailers are expected to embrace the product.**
    - **It will take up less shelf space.**

    Inferrability : 1 2 3 4 5

22. The earthquake rocked northern California last week. When Aetna adjuster Bill Schaeffer visited a retired couple in Oakland last Thursday, he found them living in a mobile home parked in front of their yard.

    - **Their house collapsed into its basement.**
    - **It was pushed about four feet off its foundation.**

    Inferrability : 1 2 3 4 5

23. J.L. Henry & Co., Miami, and a principal of the firm, Henry Otero of Miami, were jointly fined $30,000 and expelled, for alleged improper use of a customer's funds, among other things. J.L. Henry hasn't any Miami telephone listing, an operator said.

    - **Mr. Otero also couldn't be reached.**
    - **He apparently has an unpublished number.**

    Inferrability : 1 2 3 4 5

24. Medical researchers believe the transplantation of small amounts of fetal tissue into humans could help treat juvenile diabetes and some degenerative diseases.

    - **But anti-abortionists oppose such research.**
    - **They worry that the development of therapies using fetal-tissue transplants could lead to an increase in abortions.**

    Inferrability : 1 2 3 4 5

Thank you very much for filling in this questionnaire!